P. S file Copy

Drainage Correlation Research Project

INTERIM REPORT #6
August 1968

SELECTED MULTIVARIATE STATISTICAL METHODS APPLIED TO RUNOFF DATA FROM MONTANA WATERSHEDS

by

Gary L. Lewis and T. T. Williams

ENGINEERING

Research Laboratories

MONTANA STATE UNIVERSITY, BOZEMAN

Drainage Correlation Research Project

INTERIM REPORT #6
August 1968

SELECTED MULTIVARIATE STATISTICAL METHODS APPLIED TO RUNOFF DATA FROM MONTANA WATERSHEDS

by

Gary L. Lewis and T. T. Williams

Department of Civil Engineering and Engineering Mechanics

Montana State University, Bozeman

Prepared for

Montana State Highway Commission

and

U.S. Bureau of Public Roads

TABLE OF CONTENTS

TABLE OF CONTENTS (Cont.)

TABLE OF CONTENTS (Cont.)

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

A principal component analysis with varimax rotation of the principal factors was performed for watershed, storm, and runoff data from five central and eastern Montana watersheds. The analyses provided information about the relative importance of 29 independent variables to the peak discharge rates and runoff volumes produced by these variables.

Storm intensity, standard deviation of storm intensities, soil and air temperature, watershed azimuth, overland slope, watershed shape, reservoir area, and watershed area were among the most successively important variables. Correlations among some of the variables for the research watersheds were also indicated by the analyses. Principal component and rotated-factor regression equations for the runoff variables were developed, and are suggested as prediction equations for ungaged watersheds in central and eastern Montana.

Chapter I

INTRODUCTION

In order to make reasonable predictions of peak discharge rates and total runoff volumes on small watersheds, an understanding of the factors causing the runoff is needed. If the important watershed and storm variables could be properly identified and measured, the relationships between these variables and the runoff could then be readily determined.

HISTORICAL BACKGROUND

Many attempts have been made to relate peak discharge rates and total runoff volumes from a watershed to their causative factors. The relationships which have been proposed usually take the form of graphs or equations relating the rates and volumes to factors that are believed to be important. The general trend in these methods is to choose the "important" factors, obtain measurements of each, and then relate the factors to the rates and volumes produced. However, the choice of factors is usually made from experience or judgment, and the proper or improper choice of factors leads to accurate or inaccurate results. A comparison of the methods indicates that there is considerable confusion as to which factors are to be used, and which

factors are more important than others. After analyzing several methods of discharge rate and runoff volume prediction, Sharp and Biswas (1965) wrote: "Exhaustive analysis of research data from small watersheds not only failed to reveal how various factors function in producing runoff, but failed even to reveal the parameters that should be used to estimate runoff." The wide range of factors used in prediction methods seems to substantiate this. Some factors appear in more of the methods than do other factors, but none of the methods agree on which set of factors are the most important.

PURPOSE AND METHODS

The study reported herein is an attempt to determine, for five central and eastern Montana watersheds, which of 29 factors are more important in producing peak discharge rates and total runoff volumes, and the regression equations relating peak discharge rates and runoff volumes to these factors. Multivariate statistical analyses are used to investigate the relative importance of each of the independent variables to the two dependent variables, peak discharge rate and total runoff volume. As explained in a later chapter, a principal component analysis of the correlation matrix of the independent variables is performed, and this is followed by a varimax rotation of the factor

weight matrix to determine which variables are most important. Regression equations for the dependent variables are found from the important variables, using the uncorrelated principal components and the rotated factor weight matrix.

Although only a few investigators have applied multivariate methods to hydrologic data, several have recommended their use in this field. The advantages which are presented in a later chapter indicate that they provide a significant improvement over some of the other methods used. The methods are employed here because they are well suited to the large volumes of data that have been generated on the watersheds being investigated. Because of the large amounts of data involved, multivariate analyses were not practical until the advent of the electronic, digital computer. Computer techniques were used extensively in the data reduction and analyses for this study.

SCOPE

Data for this study was taken from five small central and eastern Montana watersheds currently being studied for the Drainage Correlation Research Project, by the Department of Civil Engineering and Engineering Mechanics at Montana State University. The Drainage Correlation Research Project, which was initiated in 1963, is sponsored

by the Montana State Highway Commission and the Bureau of Public Roads, and is an investigation of the frequency of peak discharge rates for small watersheds in Montana. Data collection is expected to continue until September, 1969. Fifty runoff events on the five project watersheds, having peak discharge rates greater than 10 cfs, and occurring between April, 1964, and September, 1967, were studied for the investigation reported herein.

DEFINITIONS

Because certain terms and phrases are frequently used in this paper, several definitions are presented here. The definitions are those commonly used in the literature, and some are more thoroughly discussed in later chapters.

Dependent variable - the variable to be predicted from measurements of the independent variable(s) in a regression equation (e.g., peak discharge rate).

Independent variables - the variables on which measurements are obtained and substituted into the regression equation to calculate the prediction of the dependent variable (e.g., precipitation intensity, watershed area, etc.)

Regression equation - an equation for the dependent variable, derived from several measurements of this variable and the independent variable, or variables, in a manner which indicates the relationship of the independent variables to the dependent variable. (If more than one independent variable is involved, the equation is usually termed a "multiple regression equation")

Multiple linear regression equation - a multiple regression equation in which the dependent variable is

related to a sum of the independent variables, with each of the independent variables being multiplied by a different coefficient.

Coefficient of a variable - a constant to be multiplied by the measurement of a variable in a regression equation, component, or factor.

Multivariate studies - methods of studying more than two variables when the measurements of the variables are obtained simultaneously in time or space.

Linear correlation of two variables - a statistical measure of the closeness to a straight line of the graphical plot of measurements of both variables.

Components - Principal Components - Normalized Eigenvectors - derived, uncorrelated, independent variables written as the sums of the original independent variables, if each is multiplied by a coefficient.

Factor - a component whose coefficients on the independent variables have been multiplied by a constant. The squared coefficients within a factor total to the square of the constant.

Rotated factor - a factor whose large coefficients on the independent variables have been maximized.

Variate - a component, factor, or rotated factor.

Chapter II

LITERATURE REVIEW

In recent years a great many investigators have approached the problem of peak discharge frequencies from small watersheds, and have applied a great variety of techniques in analyzing the problem. Many of the investigators have used statistical methods, and several have applied multivariate analyses to hydrologic data. Reports of a large number of these investigations were reviewed in the course of the study reported herein. To discuss all the literature which was reviewed would result in an extremely voluminous and unwieldy report. It would seem more appropriate, therefore, to confine the review herein to the results of a few of the more relevant studies which bear directly on the present investigation.

INVESTIGATIONS WITH MULTIVARIATE ANALYSES

Prior to 1950, multivariate methods were well established, but were not practical in hydrology because of the time-consuming computations. However, as the high-speed computer became more accessible to investigators, the methods were recognized as a possible means of studying the peak discharge rate and runoff volume prediction problem. For example, Wong (1963) used multivariate methods to

analyze data from 90 basins in New England. The basins ranged in area from 10 to 2000 square miles, and measurements of eleven independent variables were taken on each of these basins to determine a regression equation for the mean annual flood with a recurrence interval of 2.33 years.

Wong found that a previous "ordinary" multiple regression on five of the eleven independent variables was not satisfactory because the variables, average land slope, mean altitude, tributary channel slope, stream density, and shape of basin were, according to Horton's Laws, "multicollinear," meaning that some were linearly related to others and should not be included. After obtaining measurements on six additional variables, drainage area, main channel slope, tributary channel slope, percentage of area in ponds and lakes, length of longest watercourse, and precipitation intensity; a principal component analysis and varimax rotation were performed on the data. This resulted in two new, unrelated variables or "components." Both of these components were linear functions of all the measured variables, but the new components were not linearly related to each other. Also, the first component was found to be significantly more important than the second, and each component was found to be more highly associated with certain variables. The first had large coefficients

on variables which expressed the area and length of the drainage basin, and the second was associated with the slope and topography of the drainage basin. Both components were about equally associated with mean annual flood, indicating that both were important to the dependent variable.

Because the two components indicated that size and length were not related to slope and topography, Wong decided that these two parameters would form a good set of independent variables for a multiple regression equation. He examined the correlations of mean annual flood with the components and the eleven independent variables, and found that the length of the main stream, L, was highly correlated with both the mean annual flood and the first component; and that the average land slope, S, was similarly related to the second component and mean annual flood. These two variables were consequently chosen for a multiple regression, giving:

$$\text{Log } Q_{2.33} = -1.02 + 1.29 \text{ Log } L + 0.97 \text{ Log } S$$

for the regression equation. This equation had a coefficient of determination of 0.80, which meant that 80 per cent of the variation in the mean annual flood could be explained by only two variables instead of eleven. The previous

regression on five independent variables had the same coefficient of determination, but involved linearly related variables. The study, therefore, reduced the number of "independent" variables needed to explain the same variation in mean annual flood for New England.

Eiselstein (1967) performed a similar analysis, although he was interested in runoff volume instead of mean annual flood rate. A 350-acre watershed was divided into 17 runoff plots on which data from 30 variables was obtained over a period of four years. The variables were grouped into five categories, storm variables, antecedent moisture variables, site variables, soil description variables, and the dependent variable, runoff in inches from each plot.

Because Eiselstein was aware of the inadequacy of ordinary multiple regression to provide a good prediction equation when the independent variables are not truly independent, he performed three separate analyses to show the different results that can be obtained. An ordinary linear regression analysis gave an equation in terms of all 29 "independent" variables, and 13 regression coefficients were found to be statistically significant, i.e., non-zero. This equation accounted for 77 per cent of the variation in the runoff volume, but Eiselstein was

not satisfied with the results because the non-significant variables had high coefficients of correlation with each other and with the significant variables. Also, the test for significance was not valid because correlated variables were used. This meant that the "significant" variables were a combined measure of several variables, and the "non-significant" variables could not honestly be discarded.

To attempt to separate the combined effects of several variables to the effect of each, a principal component analysis of the correlation coefficients of all combinations of the independent variables was performed. This resulted in 29 new, independent variables, or "components," which were each linear functions of all the 29 original "independent" variables. The first of these components had high coefficients in the rainfall variables, but the remaining components could not be readily associated with specific variables. This is the reason for the varimax rotation, which rotates the components to another set of reference axes so that only high or low coefficients exist on the original variables, and a better interpretation can be made.

Before rotating the "principal components," Eiselstein computed values for each component by using the original

variable data to solve the linear equations. This resulted in a numerical value of each component for each runoff event. Because the components were truly independent, a multiple regression on these values was performed. After the regression coefficient for each component was found, 18 of the 29 coefficients exhibited significance. Also, the regression equation explained 77 per cent of the variation, which was exactly the same amount explained by the ordinary regression equation. This analysis gave a good prediction equation because truly independent variables were used. However, because each significant component was a linear function of all 29 original variables, nothing could be stated about the importance of the original variables at this point in the analysis.

To further investigate the separate variables, Eiselstein's third analysis consisted of a multiple regression on rotated components instead of the original components. An initial, orthogonal variamax rotation of the original components yielded a set of components which all had low coefficients on 12 of the original variables. Because these variables were not significant to any of the rotated components, they were deemed unimportant to the runoff and discarded. Also, only the first seven components were deemed to be important because the rest each

contributed less than one per cent to the variation of the dependent variable. This resulted in seven linear equations for 17 of the original variables, giving essentially the same information as the first 29 components and variables.

A second varimax rotation of the seven components gave seven new components which could be interpreted in terms of the 17 remaining independent variables. One of the components accounted for 51 per cent of the variation in runoff, and had high coefficients on precipitation intensity and total precipitation. The second most important component accounted for 5 per cent of the variation in runoff and had high coefficients on slope, elevation, and "aspect." The third component, accounting for four per cent of the variation in runoff, had high coefficients on surface soil properties. The other components each explained only a small portion of the variation in runoff, and could not be associated with specific variables. The important interpretation from the components was that the rainfall variables were much more important to runoff than the aspect and soil properties, because they accounted for 51 per cent of the variation in runoff. Also, the variables within each component were linearly dependent and their combined effect was independent of the combined effect of the variables of other components. This meant that a

multiple regression on the components, or on certain variables from each component, would be a regression on truly independent variables. (Wong had made this same observation in 1963 and had written his final regression equation in terms of two nearly independent variables, one from each important rotated component. Other variables were significant to the components, but the two he chose were combined measures of those in each component.)

Eiselstein chose to find a multiple regression equation for all 17 of the variables in the seven components, rather than select one variable from each component to represent the total component. After substituting data from the original variables into the component equations, a multiple regression of the values computed gave an equation for runoff in terms of the components, and hence in terms of the 17 independent variables, because the components were linear functions of the variables. The coefficients of this final equation seemed to be realistic because no obvious fallacies in the signs of the coefficients could be detected. For example, the rainfall characteristics were directly and not inversely related to runoff as is the case in some ordinary multiple regression equations (Sharp, Gibbs, Owen, and Harris, 1960; Wallis, 1965). The final coefficient of determination was 0.67 indicating

a 10 per cent loss of information as a result of reducing 29 variables to 17, and 29 components to 7.

Snyder (1961) performed a principal component regression analysis relating total December runoff from a watershed to the rainfall in October, November, and December. A previous, ordinary multiple regression equation had negative coefficients on all the independent variables indicating that runoff increased as the monthly rainfall decreased. Because this was "intuitively" inaccurate, Snyder obtained three independent components from a principal component analysis, and derived a regression equation for these components, and hence for the variables, and found that all coefficients were positive. A reduction in the coefficients of determination from 0.83 to 0.75 for the principal components solution resulted, but Snyder felt that the loss in information about the variation in runoff was justified by the intuitively correct coefficients. This study was used to illustrate the possibilities for multivariate analyses in hydrologic studies, and no rotation of the principal components was made, because the component regression equation was deemed to be satisfactory.

The three above investigations used essentially the same techniques. In all of them, the ordinary multiple regression solution was unsatisfactory, and multivariate

techniques were used to overcome this difficulty. Each used a principal component analysis to obtain truly independent variables for the regression. Also, rotation of these components was used to provide information on the importance of certain variables in two of the papers.

RECOMMENDATIONS OF THE LITERATURE

The previous section discussed some of the applications of multivariate analyses. A short presentation of several recommendations found in the literature follows. The purpose here is to present the opinions of several authorities on the use of multivariate methods with hydrologic data.

Eiselstein's investigation was undertaken as a result of a paper by Wallis (1965), which suggested the advantages of multivariate methods over ordinary multiple regression for hydrologic studies. Wallis compared the presently used methods of obtaining an equation for the dependent variable in terms of several independent variables, and made several recommendations. First, he suggested that a linear-logarithmic transformation model, similar to Wong's, be used with hydrologic data. This should be followed by a principal component regression analysis with varimax rotation of the principal components for an initial analysis

of "multifactor hydrologic problems." These recommendations were made after a study of the adequacy of the methods in obtaining a known functional relationship, the weight of a solid cylinder in terms of its density and dimensions.

Wallis (1968) presented several other suggestions for the best utilization of multivariate statistical methods in hydrologic studies. First he suggests that no more than two variables important to any factor in the rotated factor table be retained for further study. After selecting the retained variables, Wallis suggests that a complete principal component analysis with varimax rotation of the principal factors be performed on the original measurements of only these variables. A regression on these factors is then suggested for the prediction equation.

After his analysis of 90 New England basins, Wong (1963) stated that, "multivatiate methods should be more widely encouraged in geomorphic and hydrologic research" (p. 198). Eiselstein recommended that, "a principal component analysis with varimax rotation of the factor weight matrix is a suitable statistical technique for the correlation of small watershed surface characteristics with surface runoff" (p. 484). Although Snyder did not use or recommend varimax rotation, he did state that principal component

regression analyses give "logical equations" for runoff when compared to ordinary multiple regression techniques.

METHODS APPLIED TO MONTANA WATERSHEDS

Multivariate methods have not been previously applied to small watersheds in central and eastern Montana. Boner (1963) presented a report of his study of the frequency and magnitude of floods in eastern Montana for the United States Geological Survey. This report presents an ordinary multiple regression equation for mean annual flood in terms of area, stream meander length, geographical region, and elevation of small watersheds in eastern Montana. The dependent variable was found to be directly proportional to the first three variables, and inversely related to elevation. Also, the flood having a recurrence interval, I, is obtained by multiplying the mean annual flood by a factor from a "composite flood frequency curve" for that interval. Variables other than those listed were not used because measurements were not available.

Boner and Omang (1967) presented a report on the magnitude and frequency of floods from watersheds smaller than 100 square miles in area in Montana. This report gives a method of obtaining the floods with recurrence intervals of 10 and 25 years for this region. The 10-year flood is

found from a regression equation involving the area, elevation, channel slope, and mean annual runoff of a watershed, and the 25-year flood is obtained from the 10-year flood upon multiplication by an empirically determined constant.

Both of the above studies employed ordinary multiple regression techniques with only four easily determined independent variables.

The studies of Wong, Eiselstein, and Snyder used more variables than four, but the developed equations could not reasonably be used in Montana. However, the methods should be applicable to any region. None of the literature reviewed indicates that attempts have been made to determine equations for both the peak discharge rate and the runoff volume in a single analysis.

Chapter III

THEORETICAL DEVELOPMENT

In this chapter, a survey of the available methods of analyzing runoff from small watersheds using measurements of several related variables is presented, followed by a discussion of the reasons the specific methods for this analysis were chosen. A complete theoretical development of the methods chosen concludes the chapter.

POSSIBLE METHODS OF ANALYSIS

Ordinary multiple linear regression is one of the few statistical methods of simultaneously analyzing several variables to estimate one or more of them. The analysis is simply an analytical method of plotting a line, plane, or hyperplane through a multitude of data points. In general, a linear equation with an unknown intercept and slope is assumed, and the intercept and slope are calculated from the data in a manner which minimizes the squared distances from the points to the line, plane, or hyperplane. Until recently, ordinary multiple linear regression of the logarithms of the variables has probably had the most use in predicting runoff from small watersheds (Sharp, *et al*, 1960).

The greatest objection to the use of multiple regression analyses with hydrologic data is that certain basic assumptions of multiple regression theory are violated. Multiple regression assumes that there are no correlations among the independent or "predictor" variables (Thurstone, 1947). Hydrologic variables, as shown by Wong, usually violate this assumption. If ordinary multiple regression is attempted, the coefficients of the multiple regression equation for the dependent variable have sometimes been found to be "absurd" and "grossly in error" (Thurstone, 1947, p. 61).

Many regression equations for runoff have been derived using many combinations of independent variables. Snyder (1962) studied some of these equations and concluded that multiple regression analyses do not yield "logical equations" when used in hydrology. In these cases, he was referring to the coefficients associated with each of the independent variables and not necessarily with the accuracy of the prediction equation. A presentation of the mechanics of multiple linear regression was given by DuBois (1957), and Baggaley (1964).

Another statistical approach to the problem of investigating several variables was developed by Harris, et al (1961). Their purpose was to present a method of

selecting the most important independent variables, and to use only these variables in the regression equation. A Taylor series expansion was used to determine successively the most important variables by initially removing the effects of the other variables. The method is an analytic approach to "graphical curvilinear multiple regression," and eliminates the usual "shotgun" search for important variables. The method provides a statistical means of determining the important variables, but because hydrologic variables are generally correlated, as discussed earlier, attempts at multiple regression are often not successful.

Wong (1963) discusses several other possible methods of analyzing runoff in terms of several independent variables. One of these, "stepwise multiple regression," is similar to Harris' analysis because the variable which contributes most to the variation in the dependent variable is determined. Its effects are then removed, and the second most important variable is found. This process is repeated until enough variables are found to account for all or most of the variation in the dependent variable, and a multiple regression is performed on these variables. Again, the relative importance of the independent variables is indicated, but the multiple regression

assumption may be violated. Ralston (1960) presents a development of stepwise multiple regression, and outlines the procedures for programming the method for digital computers.

Another method of handling this problem is by using multivariate statistical analyses. Several of these are outlined in the next section, along with the advantages and disadvantages of each. The methods chosen for use in this investigation are indicated, and the reasons for the selection are outlined.

MULTIVARIATE STATISTICAL METHODS

One commonly used multivariate statistical analysis is known as "component analysis." There are two varieties of component analysis, known as "principal component analysis" and "centroid analysis." Both are mathematical means of obtaining new variables or "components" from the intercorrelations of the chosen independent variables. The objective in the analysis is that the components obtained will, (1) be fewer than the number of independent variables under study, (2) represent or reproduce the original variables, (3) account for all the variation in the original variables and, (4) be uncorrelated even if the original variables were highly correlated with each other. Principal

component analysis extracts the new variables, or "variates," one by one. The first component is extracted in such a manner that it reproduces a maximum amount of the information in the original data. The second component accounts for a maximum amount of the information remaining after the extraction of the first component. This process is repeated until all the components have been extracted, and all the original information in the data is reproduced by the components.

Centroid analysis has been termed a "simplified approximation of the principal components solution" (Cooley and Lohnes, 1962, p. 153), and is used to avoid the involved calculations of a principal component analysis, namely, the solution of the "characteristic equation," defined later. Kendall (1957) gives an example of the "approximate" centroid solution compared to the principal component solution of the same problem, and shows the different solutions obtained. When computer availability obviates most concern for involved computations, the principal component analysis is the better method.

"Factor analysis" is a multivariate method similar to component analysis, but differing in the general method of approaching a problem. A factor is defined as a linear equation in terms of all or some of the original variables,

but is not the same as a component. Kendall (1957) states that component analyses attempt to proceed from the data to a model, while factor analyses begin with a model and investigate its agreement with the results. In factor analyses, an investigator examines the data and speculates on how many factors or groupings of variables might be present. Guilford (1952), in a discussion on when to factor analyze, illustrated this point by writing: "The initial planning should emphasize the formation of hypotheses as to what factors are likely to be found in the selected domain and to the probable properties of such factors" (p. 36). For example, if the relationship of peak discharge rate to several variables such as soil moisture, air temperature, wind speed, watershed area, storm duration, stream channel lengths, excess precipitation intensity, soil permeability, etc., is desired, then the variables might initially be viewed as being composed of two factors or variable "groupings," one of climate, and one of watershed characteristics. The results of the factor analysis are two or more factors, and an examination of the loadings of the factors will reveal if the two-factor assumption was correct. In hydrology, factors might easily be formulated, although none of the reviewed papers employed this method. Eiselstein (1967) grouped the original independent variables into four

categories, but he did not initially state that he expected a factor for each category. His first principal component could have been termed a "rainfall" factor, but the remaining components were not readily associated with the categories. His purpose, that of principal component analyses, was to derive new, truly independent variables for a multiple linear regression. Components are independent, and the combined effect of the variables within each component is independent of the effect of other components. This means that the important variables within a component are not necessarily from the same category, as is hopefully the case with factor analyses. No initial assumptions about the outcome are made with component analyses. The data is analyzed only for uncorrelated components, and the model differs from that of a factor analysis in the above manner. If factors result, they are coincidental, but are desirable because they give information about the importance of groups of variables.

Neither a factor analysis nor a component analysis provides information about the importance of each independent variable. The coefficients on the variables are correlations of the variables with the factors or components, but they generally are relatively large in magnitude for all the variables. If one or more variables in a component

analysis has a small correlation with all the components, then the variable is not important to any of the components, and hence to the total problem. Because the first component is found in a manner which yields high correlations with all the variables, no interpretations can be made, and the components are useful only when regression of uncorrelated variables is desired.

Because a principal component analysis does not usually provide information about the importance of each independent variable, another multivariate method, "rotation of the principal components," is in general use. The components can be rotated so that each component has high coefficients on certain variables and low coefficients on other variables, allowing interpretations for the important variables, and obtaining the "simple structure" of the components (Matalas, 1967). Graphical rotation presentations are given by Fruchter (1954) and Baggaley (1964). Both authors use two-dimensional plots which show the measurements of the variables as points, and rotated components are simply lines drawn through clusters or "streaks" of points so as to maximize zero-loadings on the components. However, the graphical solutions are approximate, and different investigators might obtain different results with the same data. Kaiser (1958, 1959), derived an

analytic rotation which guarantees the same results for different investigators. This method is known as "varimax rotation," and has been utilized in the literature (Rice, 1967; Wallis, 1965; Wong, 1967; Eiselstein, 1967).

Rotation of the variates from either a component or factor analysis has no effect on the amount of information retained by the variates. Only the interpretation of the loadings is affected (Cooley and Lohnes, 1962). Kaiser's "normal" varimax rotation not only minimizes the "in-between" loadings, but it also maintains an orthogonal or perpendicular reference frame (Wallis, 1965; Kaiser, 1958). This feature is desirable if a multiple regression on the rotated components is to be performed, because perpendicularity of the components means that they are uncorrelated, and the assumptions of multiple regression are not violated. "Oblique," or non-perpendicular analytical rotations such as the "Quartimin," "Oblimin," or "Covarimin" have been developed, but do not appear satisfactory (Cooley and Lohnes, 1962).

Multivariate methods are advantageous in several aspects. Ordinary multiple regression provides good prediction results, but gives no insight into the interrelationships of the variables. Multivariate methods obviate

the effects of highly correlated "independent" variables, while ordinary multiple regression analyses do not. Interpretations of the results of multivariate analyses allow the exclusion of unimportant variables and the recognition of the more important variables.

Multivariate methods also allow the reduction of the number of variates for multiple regression. Component analyses produce exactly as many new variates as there are independent variables. However, since the extraction is done on a "priority" basis, some of the latter variates may reproduce only a small portion of the information, and may be excluded. Wong (1967), Eiselstein (1967), and Rice (1967) were all able to considerably reduce the number of variates needed to reproduce almost all of the information present.

Besides allowing the interpretation of the importance of each variable to the original data, the orthogonality of the new variates is a principal achievement. Because the data can be expressed by uncorrelated variates, then multiple regression assumptions are not violated if the regression is performed on these variates. The new independent variates reproduce the original data and are truly uncorrelated. Because of this, the correlations among the independent variables are indicated with multivariate

methods by producing uncorrelated variates.

Still another advantage of multivariate methods with correlated variables is the improvement in the final regression equation coefficients. Snyder (1962) stated that multivariate methods yield "nice" coefficients which indicate the relative importance of each independent variable to the criterion. Wallis (1965) demonstrates that principal component regression coefficients tend to be "stable" when compared to ordinary regression coefficients. In a discussion of Rice's paper, Anderson (1967) stated that the "coefficients remained very distinctly realistic with regression on principal components" (p. 6). This, and the other advantages of multivariate methods seem to provide justification for their use in hydrologic studies.

Investigators of the many multivariate methods agree that the best statistical system of analyzing hydrologic data would start with a principal component extraction, followed by a varimax rotation for interpretation of the variables, and a multiple regression on either the principal components or the rotated factors for the prediction equation (Eiselstein, 1967; Wallis, 1965; Wong, 1967; Anderson, 1967). Baggaley (1964), however, states that good rotation results are not likely unless more than 20 variables are involved. Because 29 independent variables were available

for the analysis reported herein, multivariate methods were indicated for the present study.

DEVELOPMENT OF METHODS

The methods presented in this section were developed by Kendall (1957) and Kaiser (1958), and most of their equations and derivations are repeated herein. Graphical interpretations of the methods are included in the Appendix. Some of the steps in Kendall's and Kaiser's derivations were omitted in their reports, and are included in the present development because of their importance.

Model

In all studies which involve regression as a final analysis, a model for the regression must be assumed. The variables may be powered, multiplied, divided, or added. In general, multiple regression analyses provide the coefficients for linearly additive independent variables. This equation usually has the form

$$Y = B_0 X_0 + B_1 X_1 + \ldots + B_p X_p \qquad (1)$$

where Y is the dependent variable or criterion, X_i is the $i^{\underline{th}}$ independent variable, and B_i is the desired regression coefficient. In general, X_0 is unity, and the equation is that of a "hyperplane" in (p-1) dimensions.

Hydrologic variables are generally multiplicative in nature (Wallis, 1965), and plotting one variable against another on logarithmic paper usually approximates a linear relationship. For this reason, hydrologists often assume that the dependent variable will be related to the p independent variables in the form

$$Y = B_0 X_1^{B_1} X_2^{B_2} X_3^{B_3} \ldots X_p^{B_p} \qquad (2)$$

Upon taking logarithms of this equation, a linear equation having the form

$$\text{Log } Y + \text{Log } B_0 + B_1 \text{ Log } X_1 + B_2 \text{ Log } X_2 + \ldots + B_p \text{ Log } X_p \qquad (3)$$

is obtained. If the logarithms of the variables are used as variables, then a linear multiple regression may be performed that will provide the coefficients for Equation (3). Equation (3) is the model used herein to represent the equations for peak discharge rate and total runoff volume. The logarithm to the base 10 of the measured variable is represented hereafter by X_k.

Principal Component Analysis Theory

The attempt in a principal component analysis is to find a set of linear equations which reproduce the information present in a set of measurements of several independent variables. Hopefully, the number of linear

equations required will be less than the number of variables rather than several. These new variables are referred to as "principal components," or simply as "components" in the literature. The linear form of the components, for the $i^{\underline{th}}$ value of the $j^{\underline{th}}$ component, is

$$V_{ji} = \sum_{k=1}^{p} l_{kj} \, x_{ki} \tag{4}$$

where x_{ki} is the $i^{\underline{th}}$ measurement of the standardized form of the variable X_k, or

$$x_{ki} = \frac{X_{ki} - \overline{X_k}}{s_{X_k}} \tag{5}$$

and l_{kj} is an undetermined coefficient for the $k^{\underline{th}}$ variable and the $j^{\underline{th}}$ component. The other terms, $\overline{X_k}$ and s_X are the mean and standard deviation, respectively, of the $k^{\underline{th}}$ variable. The purpose of standardization is to give the standardized variables a mean of zero, and a variance equal to unity. For n observations of the standardized variable x_k, the mean is

$$\frac{1}{n} \sum_{i=1}^{n} x_{ki} = 0; \quad k = 1, 2, \ldots, p \tag{6}$$

and the variance is

$$\frac{1}{n} \sum_{i=1}^{n} x_{ki}^{2} = 1; \quad k = 1, 2, \ldots, p \tag{7}$$

Standardization does not change the importance of the variables, and the x_k terms are now dimensionless because the standard deviation has the dimensions of the variable. Therefore, if the variables are standardized before the analyses are begun, the effects of the different dimensions of each variable are not present, and better interpretations are possible.

To solve for the l_{kj} values in Equation (4), two conditions must be satisfied. First, it is desired that the components be statistically uncorrelated (truly independent) so that a regression of the components may be performed without violating the independence assumption. This means that the "correlation coefficient" for any two components must equal zero. The correlation coefficient is a statistical measure of the degree of linear association between two variables. The value of the coefficient ranges between plus one and minus one, depending on the increase or decrease of one variable, respectively, as the other variable is increased. A correlation of plus or minus one would mean that the variables were perfectly related, and a line plotted on a two-dimensional graph would intersect every point. For any two components to be uncorrelated,

$$r_{bc} = 0 \tag{8}$$

must be satisfied, where r is the correlation coefficient, and b and c are the subscripts of the respective components.

The second condition for solution for the l_{kj} values in Equation (4) is that the first component must reproduce a maximum amount of the information in the original measurements of the variables. A statistical measure of the information contained by a component is the variance of the component. Because the variables are standardized, they each have a variance of unity, and the total variance is equal to p, because there are p independent variables. If the first few components are to reproduce the information in the measurements of the variables, then the sum of the variances of the components must also equal p. Therefore, the component which has a maximum variance is desired. To determine the component, it is necessary to write the equation for the variance of any component in terms of the unknown l_{kj} values, and then equate the partial differentials of this equation, with respect to l_{kj}, to zero. This procedure gives the l_{kj} values which maximize or minimize the variance of the first component.

Kendall (1957, p. 14) takes the partial differentials of the equation for the variance of a component and finds the solution which minimizes or maximizes the variance.

If r_{jk} is the correlation of any two variables defined by

$$r_{jk} = \frac{1}{n} \sum_{i=1}^{n} x_{ji} \, x_{ki} \tag{9}$$

and if L is an "undetermined multiplier" (see Appendix A for a graphical derivation), then the criterion for the extremum is the determinant (Kendall, p. 15)

$$\begin{vmatrix} (1-L) & r_{12} & r_{13} & \cdot & \cdot & \cdot & r_{1p} \\ r_{21} & (1-L) & r_{23} & \cdot & \cdot & \cdot & r_{2p} \\ \cdot & & & & & & \cdot \\ \cdot & & & & & & \cdot \\ r_{p1} & r_{p2} & r_{p3} & \cdot & \cdot & \cdot & (1-L) \end{vmatrix} = 0 \tag{10}$$

If I is defined as an identity matrix, then the determinant can be written

$$\left| R - LI \right| = 0 \tag{11}$$

in matrix notation. The matrix of correlation coefficients of the standardized independent variables, R, is easily computed by applying Equation (9) to all combinations of variables, and L is the only unknown.

Equation (11) is referred to as the "characteristic" equation of the correlation matrix. Several solutions to this equation are available (Cooley and Lohnes, 1962; Ralston, 1960) which yield the L's and a set of l_{kj} values

for each L. Also, there are generally p values of L which satisfy Equation (11). These values are called "eigenvalues," "latent roots," "characteristic roots," "proper values," or "characteristic values" by various authors. The term "eigenvalues" is used herein.

The set of l_{kj} values corresponding to any L is known as the "eigenvector" for that eigenvalue. If each l_{kj} value in an eigenvector is divided by the square root of the sum of squared l_{kj} values, then a "normalized eigenvector" results. Some authors (Kendall, 1957) call the normalized eigenvectors "principal components," and this notation is used herein. Other authors (Eiselstein, 1967; Harman, 1967) multiply the normalized eigenvector "loadings" by the square root of the respective eigenvalue and call this new vector the principal component, or "principal factor." In the present analysis, a principal component is identical to a normalized eigenvector, and a factor, whenever cited, is meant to be defined as

$$A_j = \sqrt{L_j}\, V_j \tag{12}$$

where the normalized eigenvector, V_j, is defined by Equation (4), and L_j is the eigenvalue corresponding to the $j^{\underline{th}}$ eigenvector. The l_{kj} values in Equation (4) are coefficients of normalized eigenvectors.

Kendall (1957, p. 15) shows that the variance of a component is

$$\text{Variance } (V_j) = \sum_{i=1}^{n} \left(\sum_{k=1}^{p} l_{kj}\, x_{ki}\right)^2 \qquad (13)$$

which is also the component's eigenvalue. This means that the component having the largest eigenvalue reproduces a maximum amount of the information contained in the independent variable correlations. The component having the second largest eigenvalue reproduces the next largest amount of the information, etc., for all p components.

The sum of the eigenvalues for a "symmetric" matrix, such as the correlation matrix used herein, equals the sum of the principal diagonal entries (Hotelling, 1933, p. 429). Because a principal component analysis uses the self-correlations of the variables in this diagonal, then this sum is equal to p, because p independent variables are used. Also, as shown earlier, the total variance in the standardized independent variables is equal to p, and the eigenvalues are the variances of each component. The total variance of the components is equal to the total variance present, and the components therefore contain all the information in the correlation matrix, which in turn contained the information in the measurements of the data. Also, the variance of each component is the portion of the

total information reproduced by the component. The per cent of the total variance "accounted" for by any component is therefore given by

$$\% \text{ Variance} = \frac{L_j}{p} (100) \tag{14}$$

Earlier, it was stated that as few components as possible were desired to account for the total variance in the independent variables. Because each successive component accounts for a maximum amount of the remaining variance in the original variables, then some of the later components can be expected to have small eigenvalues. Originally, there are p components accounting for 100 per cent of the variance. If, for example, 30 independent variables were under investigation, then 30 components would be computed. However, if 10 of the components accounted for 90 per cent of the variance, then 20 components could be dropped if the 10 per cent loss in information was justified by the reduction in variables for regression. The regression equation would contain the 30 original variables, but would be performed on the 10 uncorrelated components and the dependent variable, and not on the 30 independent variables. The resulting equation should be a better prediction equation than one obtained from ordinary multiple linear regression with all 30 variables,

because the regression assumption of independence would not be violated.

No analytical method of determining which components are important could be found in the literature. Only subjective observations of the values have been used. Kendall (1957) uses only those components whose eigenvalues are significantly larger than others. In one example, he performed a principal component analysis on five variables and arrived at five eigenvalues: 3.2470, 1.2753, 0.3859, 0.0700, and 0.0218, totaling to 5.0000. Because of the large difference between the third and fourth, he used only the first three components in his regression equation for beer consumption, accounting for 98 per cent of the original variance. Cooley and Lohnes (1962, p. 160) state that if unities were used in the principal diagonal of the correlation matrix, then only those components whose eigenvalues are greater than unity should be used in future analyses. Wallis (1965) recommends the use of as many new variates as are needed to account for as high as 99.5 per cent of the variance, thereby retaining most of the original variance. Other authors are aware of this problem, but are not specific in their methods of solution. Anderson (1967) agrees with Kendall's method of observing a "large change" in the eigenvalues. Hotelling (1933, p. 421)

suggests that the components, "whose contributions to the total variance are small" be neglected. Eiselstein (1967) neglected those components (dimensions) whose contributions to the total variance were less than one per cent. Because of the relatively small use of principal component analyses to date, no mathematical method for determining which variates to use has yet been established. According to the literature, the investigator must analyze his own results for the solution to this problem.

In the above paragraphs, it has only been stated that the components derived are uncorrelated, or that Equation (8) applies. Kendall (1957, p. 16) proves that the correlation of any two components is zero, and that all the components are perpendicular, forming a "p-dimensional reference frame." For this reason, principal components are sometimes referred to as "principal axes."

The principal components provide a means of deriving uncorrelated variates for multiple regression. Graphically, they are reference axes drawn through the measurements of the independent variables so that a maximum amount of the information contained by the measurements is explained by the components. However, no information about the relative importance of the original variables is presently available, and this is the purpose of the

next section, which presents varimax rotation theroy.

Varimax Rotation Theory

Varimax rotation of normalized factors from a factor analysis was originated by Kaiser (1958). Kaiser (1959) developed a computer program to perform the rotation. A graphical explanation of factor rotation was not presented by Kaiser, and is therefore presented in Appendix B.

The set of axes through the data points which gives the best interpretation of the contribution of each independent variable to the information is accomplished with varimax rotation by the rotation of the principal axes discussed previously. If the b_{kj} values are the desired coefficients for the rotated principal factors, and if the a_{kj} values are the coefficients from Equation (12), then the variance of the kth rotated factor is given by Kaiser (1959) as

$$\text{Variance } (A_k) = \sum_{j=1}^{s} \frac{p \sum_{j=1}^{s} b_{kj}^4 - (\sum_{j=1}^{s} b_{kj}^2)^2}{p^2} \qquad (15)$$

which he calls the "varimax criterion" for maximum interpretation (see Appendix B for a graphical interpretation).

Because the a_{kj} values are known, and because the b_{kj} values are geometric functions of the a_{kj} values and the

angle of rotation of the principal axes, θ, which maximizes the criterion, then the angle is the only unknown, and

$$b_{k1} = a_{k1} \cos \theta + a_{k2} \sin \theta \tag{16a}$$

$$b_{k2} = a_{k2} \cos \theta - a_{k1} \sin \theta \tag{16b}$$

for the first and second factors (Kaiser, 1959). The desired angle in the plane of these factors can be obtained by substituting b_{kj} values from Equations (16) into Equation (15), differentiating with respect to θ, and solving for the angle which makes the differential equal to zero. After the angle is found, the values of b_{k1} and b_{k2} can be computed from Equations (16), giving the loadings of the rotated factors. If these single-plane rotations are made for all combinations of two factors, then the complete set of b_{kj} values can be found. Substitution of these values into Equation (15) will give a value of the criterion which should be a maximum, and the b_{kj} values are the coefficients relating the rotated factors to the independent variables. The conditions for a maximum of Equation (15) are found by Kaiser who takes the second derivative of Equation (15) with respect to θ, and solves for the limits on θ which make the derivative negative. If θ for any single plane rotation is not within these limits, then the angle is equated to the nearest limit,

and a second rotation is made from this point. This process will converge to the solution for Equation (15) which is a maximum (Kaiser, 1959), giving the desired b_{kj} values.

Because the perpendicularity of the factors is maintained by Kaiser's normal varimax rotation, the final factors are uncorrelated and provide independent variates for regression. The principal components are also perpendicular, but give no information about the variables because the coefficients are not derived for this purpose. If certain variables are not highly correlated with any of the rotated factors, then they do not contribute to the information contained by the factors, and they may be deemed unimportant to the information. Also, each rotated factor accounts for a percentage of the total information, and if some factors are not important, then any variables correlated only with these factors are also unimportant. Certain independent variables may therefore be regarded as unimportant and excluded from the regression analysis of the rotated factors.

Multiple Regression of Principal Components and Rotated Factors

A principal component analysis could be performed with no rotation for interpretation of the variables, if the independent variables used were all known to be

important. A multiple regression of the principal components would give an equation for the dependent variable in terms of all the independent variables with no regression assumption violations. If this was the only purpose of the analysis, then one of two methods could be used to obtain the regression coefficients of the independent variables.

One method of obtaining the coefficients for the principal components is to solve the equations of the components using all the measurements of the original independent variables, giving n sets of values for the s components, one set for each measurement of the dependent variable. The log-transformed data from one measurement of all the variables is substituted into Equation (4) for all s components, giving one set of the new independent variates for the respective measurement of the dependent variable. This is repeated for all observations, and an ordinary linear multiple regression of the dependent variable and the values for the components is performed. This gives the regression constant and coefficients for the equation of the $i^{\underline{th}}$ observation

$$y_i = F_0 + \sum_{j=1}^{s} F_j V_{ji} \tag{17}$$

where the F_j values are the regression coefficients of the

components, and y is the standardized dependent variable. Substituting Equation (4) for V_{ji} gives the regression equation in terms of the original standardized independent variables

$$y_i = F_0 + \sum_{j=1}^{s} F_j \sum_{k=1}^{p} l_{kj} x_{ki} \tag{18}$$

If y and x_k are the standardized form of the logarithms of the original variables, then

$$x_{ki} = \frac{\text{Log } X_{ki} - \overline{\text{Log } X_k}}{s_{\text{Log } X_k}} \tag{19}$$

where X_{ki} is the $i^{\underline{th}}$ measurement of the variable X_k. Substituting this and a similar equation for y into Equation (18) gives the regression equation in terms of the measured variables

$$\text{Log } Y = \sum_{k=1}^{p} B_k \text{ Log } X_k + C \tag{20}$$

Equation (20) is valid only if

$$B_k = s_{\text{Log } Y} \sum_{j=1}^{s} (F_j \frac{l_{kj}}{s_{\text{Log } X_k}})$$

and if

$$C = s_{\text{Log } Y} F_0 + \overline{\text{Log } Y} - s_{\text{Log } Y} \sum_{j=1}^{s} \sum_{k=1}^{p} (F_j \frac{l_{kj}}{s_{\text{Log } X_k}} \overline{\text{Log } X_k})$$

Equation (20) reduces to the form desired, given by Equation (2) as

$$Y = B_0 X_1^{B_1} X_2^{B_2} X_3^{B_3} \ldots X_p^{B_p} \tag{2}$$

if

$$B_0 = \text{Antilog}(C)$$

Because the original data must be used twice for the computation of the coefficients in Equation (2); once for the correlation matrix computation, and once in the regression analysis, a better method of obtaining the coefficients with fewer computations may be used. Kendall (1957) shows that the regression coefficients for the principal components (normalized eigenvectors) can be computed from the eignevalues by using the equation

$$F_j = \frac{\sum_{k=1}^{p} l_{kj} \, r_{ky}}{L_j} \tag{21}$$

which can be derived using multiple regression theory. The values of r_{ky} are the correlation coefficients of the kth standardized independent variable with the dependent variable y, and are known from the correlation analysis. The other terms are defined earlier, and the F_j values can be computed without returning to the original data.

Once the F_j values are found, Equation (2) gives the desired relationship.

A simplified method for multiple regression of rotated factors could not be found in the literature, but a method exactly the same as the method indicated by Equation (17) for principal component regression may be employed. Because the rotated factors are given by the equation (see Appendix B)

$$A_{ji}' = \sum_{k=1}^{p} c_{kj} x_{ki} \tag{22}$$

where the prime indicates a rotated factor, and the c_{kj} values are the factor loadings, then Equations (17) through (20) may be applied to the rotated factors, giving

$$y_i = G_0 + \sum_{j=1}^{s} G_j A_{ji}' \tag{23a}$$

$$B_k = s_{\text{Log } Y} \sum_{j=1}^{s} \left(G_j \frac{c_{kj}}{s_{\text{Log } X_k}}\right) \tag{23b}$$

$$C = s_{\text{Log } Y} G_0 + \overline{\text{Log } Y} - s_{\text{Log } Y} \sum_{j=1}^{s} \sum_{k=1}^{p} \left(G_j \frac{c_{kj}}{s_{\text{Log } X_k}} \overline{\text{Log } X_k}\right)$$

where the G_j values are from an ordinary linear multiple regression of the rotated factors and dependent variable, using only those independent variables deemed important.

Unlike the component regression, an equation for the dependent variable in terms of the important independent variables is obtained with Equation (23a). This equation, or the one obtained from a principal component regression, should give reasonable predictions of the dependent variable, and should exhibit sensible signs and magnitudes for the coefficients of the variables.

Summary

The methods reviewed herein begin with the computations of the correlations of all the standardized variables, dependent and independent. This is followed by the principal component analysis of the correlations among the independent variables. The components are then converted to factors which are rotated, allowing the interpretation of the importance of the independent variables to the information contained in the original measurements. If the measurements of all the independent variables may be obtained, then the principal component regression equation provides the best prediction. Otherwise, an adequate regression equation from the rotated factors and important independent variables may be used.

Chapter IV

ANALYSIS OF DATA

The procedures developed in Chapter III were applied to the watershed data in the order discussed in that chapter. The methods used and results obtained in recording the data, selecting the variables, computing the values of the variables, and analyzing the values are presented in this chapter.

DATA RECORDED

The measurements of 29 independent variables and two dependent variables for 50 runoff events were obtained from continuous data taken from five central and eastern Montana watersheds. The watersheds were selected and instrumented in 1963 (Williams, 1965). Residents living on or near the watersheds were consulted, and a system for instrumentation of the watersheds was established by those residents who agreed to service and maintain the instruments. Table I gives the watershed names, locations, sizes, and instrumentation numbers. The weather stations consisted of instruments which measured and recorded, at half-hour intervals, the following information: soil moisture (per cent on a dry weight basis) and temperature (deg F) at 3, 9, and 18 inches below ground surface; air

Table I. Instruments Located on Watersheds

Watershed	Lone Man Coulee	Bacon Creek	Hump Creek	Duck Creek	East Fork Duck Creek
County	Pondera	Wheat-land	Sweet-grass	McCone & Praire	Praire
Area (sq mi)	14.10	17.97	7.61	53.79	13.67
Runoff Events	20	4	3	10	13
Non-recording Rain Gages	2	3	1	4	4
Recording Rain Gages	2	2	2	3	2
Snow Courses	3	3	3	3	3
Weather Stations	2	2	1	3	3
Wind Stations*	1	1	1	1	1
Water-stage Recorders	1	1	1	2	1

*Some weather stations did not have instruments for measuring wind speed and direction.

(The instrument stations for East Fork Duck Creek were the same as for Duck Creek).

temperature (deg F) at 4 and 10 feet above the surface; and wind speed (mph) and direction at 10 feet above the surface. United States Geological Survey (USGS)

Quadrangle maps and some field slope measurements provided the measurements of the watershed topographic variables. The U.S. Soil Conservation Service (SCS) performed surveys and prepared maps showing the types and proportions of the soils on the watersheds. Measurements of infiltration rates for most of the soils were made by project personnel during the summer of 1966. The USGS obtained cross-section and velocity measurements near the water-stage recorders for some of the runoff events, yielding data for the construction of stage-discharge curves. Aerial photographs of the watersheds provided information about the land use and snow coverage.

Data from the continuous water-stage recorders were reduced to hydrographs relating discharge rate to time, using the stage-discharge relationships for each recorder. These hydrographs provided the criterion for selection of the runoff events to be used. Whenever the peak discharge rate was larger than 10 cfs, a starting and an ending time of discharge were established, giving the runoff hydrograph for that event. Generally, there was no base flow before the event, and the hydrograph was terminated when the discharge rate reduced to zero flow. This procedure established the 50 peak discharge rates, and the measurements of the second dependent variable, total runoff, were

determined from the area under each hydrograph.

SELECTION OF INDEPENDENT VARIABLES

Measurements of 29 independent variables for each runoff event were computed. The variables are presented in Table II, and this is followed by the reasons for the

Table II. Independent Variables Studied

Var. No.	Var. Name	Watershed and Storm Variable Definitions & Units of Measurement
1	A	Area (sq mi)
2	SHP	Shape (dimensionless)
3	AZ	Azimuth (deg)
4	ELEV	Elevation (ft)
5	GNDS	Ground slope (ft/ft)
6	GNDL	Overland distance of flow (mi)
7	FREQ	Stream frequency (1/sq mi)
8	L	Main channel meander length (mi)
9	S	Main channel straight-line slope (ft/ft)
10	USE	Land use ratio (dimensionless)
11	INFR	Soil infiltration rate (in./hr)
12	POND	Per cent ponds and reservoirs (%)
13	I	Precipitation intensity (in./hr)
14	ISD	Standard deviation of intensities (in./hr)
15	D	Duration of storm (hr)
16	TDF	Time distribution of precipitation (hr)
17	TPCP	Total precipitation (in.)
18	API	14-day antecedent precipitation index (in.)
19	SOLM	Soil moisture (%)
20	WDIR	Wind direction (dimensionless)
21	WEEK	Week of the year (dimensionless)
22	AIRT	Mean air temperature (deg F)
23	ATSD	Standard deviation of air temperatures (deg F)
24	WVEL	Mean wind velocity (mph)
25	WVSD	Standard deviation of wind velocities (mph)
26	SOLT	Mean soil temperature (deg F)
27	STSD	Standard deviation of soil temperatures (deg F)
28	DEGD	Degree-Days (deg F)
29	SWEQ	Snow-water equivalent in snowpack (in.)

selection of these particular variables. Detailed descriptions of the variables are listed in Appendix C.

A few comments concerning the detailed descriptions of the variables should be made before proceeding with the discussion of the analysis. The Thiessen areas for the various types of "weighting" were determined from repeated planimetering of reduced SCS soil maps of the watersheds. The area attributed to each weather station, snow course, or precipitation station was determined by drawing perpendicular bisectors of the lines connecting the stations. The bisectors were intersected with each other and with the watershed boundaries, forming the polygon for each station.

Because contoured topographic maps with small contour intervals were not available for the Duck Creek and East Fork of Duck Creek watersheds, nor for the Bacon Creek watershed, transit-stadia surveys of the main channels of Duck Creek and East Fork of Duck Creek were performed to establish the vertical properties of these watershed channels. The slopes obtained from the survey varied by less than 10 per cent from the slopes obtained from the contoured USGS maps. A careful analysis of these maps will therefore yield satisfactory values for the slope variables, GNDS and S. For this reason, a survey of the

Bacon Creek watershed was not conducted, and a similar large-interval map was used.

The variables measuring ground slope and overland distance of flow were measured in the manner chosen because of the difficulty of other methods of representing these parameters. Some investigators use the method employed here, while others measure a random number of distances and slopes from ridges to channels, and average these. Because the steepest portions of watersheds are generally in the upper reaches, the method employed attempts to establish the steepest and shortest path for the overland flow.

Several methods of measuring the population of stream segments on a watershed can be found in the literature (Chow, 1964). The method chosen represents the number of segments per unit area of the watershed, and therefore has a physical meaning. A large value for this variable would indicate that the watershed exhibits a large number of branching streams and tributaries.

Because aerial photographs were not available for each year of data recorded, the land-use variable remained constant for each runoff event. Other forms for this variable have been suggested, but the available data was not sufficient to warrant the use of any of them. The

index chosen increases with the vegetation on the watershed, and therefore has a physical interpretation.

The infiltration variable was chosen to measure an average infiltration rate for the different soil types. A single number for each watershed is not altogether reasonable if many soil types are present, but the agreement of the average infiltration rates on each of the three SCS soil types with published data was encouraging. Table III compares the measured values with the suggested ranges from the literature.

Table III. Measured and Published SCS Infiltration Rates

SCS Soil Type	Avg. of 5 Wshds.	Published Range*
B	0.129	0.15 to 0.30
C	0.094	0.05 to 0.15
D	0.040	0.00 to 0.05

*From Chow (1964) p. 12-26.

The pondage variable was not easily obtained, and may not be indicative of the storage which takes place during a runoff event. The purpose of using this variable was to indicate the possible storage capacity of the watersheds. Although the depths of the reservoirs are not included in the variable, the surface area is frequently found to be related to depth, and both may

be measured in the same variable.

The remaining variables were chosen because of their frequent use in similar studies. The determination of values of all the variables was time-consuming because computations of some of the variables required analyses of large amounts of data. The development of good prediction equations usually requires easily-obtained variables, and this fact was also considered when choosing the variables. Table IV

Table IV. Summary of Watershed Characteristics

Variable	Lone Man Coulee	Bacon Creek	Hump Creek	Duck Creek	E.F.Duck Creek
A (sq mi)	14.1	17.97	7.61	53.79	13.67
SHP (mi/mi)	2.448	4.332	2.569	2.721	3.684
AZ	258	289	177	164	138
ELEV (ft)	3916	4578	4144	2853	2862
GNDS (ft/ft)	0.0194	0.0162	0.0650	0.0596	0.0763
GNDL (mi)	0.878	0.516	0.239	0.820	0.380
FREQ (1/sq mi)	5.46	26.66	24.18	14.05	15.80
L (mi)	8.446	12.895	6.444	23.614	13.051
S (ft/ft)	0.00639	0.01182	0.02412	0.00596	0.00692
USE (%)	1.35	75.18	45.64	9.67	10.51
INFR (in/hr)	0.09481	0.08824	0.10155	0.10022	0.09689
POND (%)	0.02352	0.00915	0.00000	0.00614	0.02887

summarizes the physiographic variables, which were assumed to remain constant for each watershed. The values of the storm variables are not tabulated herein but are available in the Department of Civil Engineering and Engineering Mechanics, Montana State University. The means and standard deviations of the storm and physiographic variables are listed in Appendix H for the raw and log-transformed variables.

TREATMENT OF MISSING DATA

Operational and mechanical problems with the instruments during the times of interest caused some missing values for the storm variables. At least one recording precipitation station was operational during each of the storms, and hence no information was lost for the precipitation variables. Thiessen areas for only those stations which were recording were used whenever some stations were non-operative. The greatest losses of complete data occurred from the weather stations. Whenever data was missing from all the weather stations on a watershed, the average of the variable computed for the same time period over the other years of data was used. The one exception to this was the variable for wind direction. Here, the time difference of appearance of precipitation

on the recording precipitation stations was used to estimate the direction of movement of the storm, and therefore the probable wind direction.

The University computers were used extensively in the reduction of the data for the value of each storm variable for all the storms. Most of the programs written were elementary and are not included with this report. One program for computation of variables 13, 14, 15, 16, and 17 in Table II is rather involved, and is included in Appendix D. The standard Weather Bureau format for recording and non-recording precipitation data was used in punching the project data, but operations on this format were found to be difficult. The included program could probably have been much shorter if another input format had been used.

ANALYSIS OF 31 VARIABLES

Correlations

After the values of the 29 independent variables and 2 dependent variables were computed for the 50 runoff events, the data were transformed by taking the base-10 logarithms of all the values. This was done in order to use Equation (3) as the model for the prediction equation. Zero measurements of any of the variables were set equal to 0.00001 before transformation, in order to compute

logarithms with the computer. The correlation coefficients of each of the transformed variables with all other variables were then computed using a modified version of a correlation computer program from Cooley and Lohnes (Appendix E). The means and standard deviations of all log-transformed variables and of the original variables were also computed, and these are listed in Appendix H. Computations for the eigenvectors and their respective eigenvalues required the correlations between the standardized, transformed variables (Equations 9 and 10). However, since these correlations are identical to the values obtained for the log-transformed variables, as stated earlier, standardization was not employed at this point of the analysis.

Principal Component Analysis

The correlation coefficients for the 29 independent variables were entered into a modified version of a principal component computer program given by Cooley and Lohnes (Appendix F). This program computed the eigenvalues and normalized eigenvectors of the correlation matrix, giving 29 principal components and their respective variances. Table V lists the variance, percentage of total variance, and accumulated percentage of the first 18 components. Equation (14) was used in computing the per

Table V. Properties of the First 18 Components

Component Number	Variance (Eigenvalue)	Per Cent of Total Variance	Accumulated Per Cent
1	9.625	33.19	33.19
2	4.772	16.46	49.65
3	3.778	13.03	62.68
4	2.323	8.01	70.69
5	1.730	5.07	76.66
6	1.385	4.77	81.43
7	1.025	3.53	84.96
8	0.961	3.31	88.27
9	0.806	2.78	91.05
10	0.684	2.36	93.41
11	0.430	1.48	94.89
12	0.410	1.41	96.30
13	0.278	0.96	97.26
14	0.214	0.74	98.00
15	0.180	0.62	98.62
16	0.140	0.48	99.10
17	0.114	0.39	99.49
18	0.072	0.25	99.74

cent of the total variance accounted for by each component, where p was the number of independent variables. The other eleven components contributed about equally to the small per cent remaining, and were considered unimportant.

The criteria of Eiselstein, Cooley and Lohnes, and Kendall were combined in determining which components to retain for further analyses. If only those components whose eigenvalues are greater than unity (Cooley and Lohnes) were used, then only about 85 per cent of the information in the measurements would be retained by seven components. There were no obvious "large" changes in the

values (Kendall), and twelve components would have to be retained if those components having less than one per cent of the variance were discarded. The author decided to retain the first ten components, losing about 7 per cent of the information.

Table VI shows that none of the 29 independent

Table VI. Reduced Loadings for the First 10 Components

Variable	Component 1	2	3	4	5	6	7	8	9	10
A	.11	-.16	.30	.01	.25	-.39	-.29	-.01	-.08	.01
SHP	.16	-.08	-.15	-.46	-.24	.00	-.03	-.08	.13	-.06
AZ	-.19	.33	-.06	-.06	.07	-.19	.10	-.00	.12	.12
ELEV	-.16	.36	-.15	-.05	.04	-.14	.05	-.02	.09	.15
GNDS	.20	-.32	-.10	.10	-.02	.15	-.07	.00	-.14	-.10
GNDL	-.15	.09	.40	.08	.14	-.20	.15	.04	.04	.03
FREQ	.23	-.12	-.27	-.20	.04	-.18	.10	-.08	-.04	.02
L	.18	-.25	.21	-.12	.12	-.28	.21	-.03	-.05	-.04
S	.06	.14	-.46	.03	.11	-.11	.04	-.05	-.07	.13
USE	.21	-.06	-.29	-.22	.06	-.24	.14	-.08	-.03	.05
INFR	.10	-.26	-.01	.44	.20	.08	-.11	.05	-.26	-.04
POND	-.07	-.05	.37	-.29	-.27	.14	.09	.01	.20	-.13
I	.26	.19	.06	-.02	.00	.04	.14	.20	-.01	.05
ISD	.27	.18	.08	.00	.03	.05	.11	.17	.00	-.09
D	-.18	-.15	-.02	-.13	.36	.12	.07	-.31	.16	-.05
TDF	.06	-.03	.08	.09	.02	-.50	-.63	-.28	.11	-.16
TPCP	.19	.16	.03	-.30	.36	.30	.16	-.18	.15	-.05
API	.19	.04	.06	.11	.23	.29	-.20	-.33	.19	-.18
SOLM	-.09	.23	-.14	.00	.44	-.09	-.25	.16	.08	-.11
WDIR	.08	-.22	.10	-.09	.16	.05	-.35	.20	.05	.77
WEEK	.27	.17	.10	.03	.01	.03	-.06	-.01	.19	-.01
AIRT	.22	.21	.11	-.04	-.07	.07	-.27	.12	-.21	.06
ATSD	-.11	-.12	-.07	-.18	.30	.09	-.12	.61	.07	-.25
WVEL	-.17	.05	.14	-.25	.15	.23	-.01	-.35	-.42	.26
WVSD	-.13	.12	.06	-.34	.12	-.02	-.16	.04	-.58	-.26
SOLT	.28	.18	.10	.03	-.05	.00	-.03	.03	-.05	.03
STSD	.18	-.13	.09	-.34	.21	.03	-.03	.05	.21	-.00
DEGD	.26	.16	.07	.02	.00	-.05	-.05	-.02	-.25	-.12
SWEQ	-.25	-.20	-.11	-.01	-.02	-.03	-.03	.05	.03	-.12

variables had small correlations with all 10 of the important components. The reduced values in the rows of the table are the correlations of the variables with the components. The numbered components correspond to the first ten components of Table V, which gives the respective variance of each. Because no information about the independent variables could be obtained at this point, the components were reduced to factors via Equation (12), in preparation for rotation.

Varimax Rotation of Principal Factors

An initial varimax rotation of the ten factors was performed using a modified version of a computer program listed by Cooley and Lohnes (Appendix G). The program was modified to read either components and eigenvalues, or factors. Computations for the variance of each rotated factor were not initially made by the program, and statements for the computations and output were included. The program was tested with data from both Eiselstein and Harman, and the results agreed very well with those obtained by both investigators.

The ten factors were rotated by the program, giving the new factors shown in Table VII. The reduced variance, per cent of original variance, and accumulated percentage

Table VII. Properties of the 10 Rotated Factors

Factor	1	2	3	4	5	6	7	8	9	10
Variance	7.3	4.3	3.8	2.8	2.4	1.6	1.4	1.2	1.2	1.1
Per Cent	25.2	14.7	13.1	9.5	8.3	5.5	5.2	4.2	4.0	3.7
Accum. %	25.2	39.9	53.0	62.5	70.8	76.3	81.5	85.7	89.7	93.4
A	.14	-.20	.08	.11	-.94	.02	.02	-.02	-.11	.09
SHP	.18	-.09	-.94	.22	.08	-.02	-.02	-.02	.01	.04
AZ	-.08	.91	.35	-.03	.04	-.14	-.06	.00	.02	-.11
ELEV	-.06	.87	.14	-.31	.30	-.09	-.05	.01	.04	-.14
GNDS	.09	-.90	-.35	-.11	-.07	.16	.06	-.00	-.02	.11
GNDL	-.09	.40	.68	.35	-.48	-.11	-.03	-.00	-.04	-.00
FREQ	.24	-.28	-.82	-.37	-.17	.09	.02	-.02	-.03	.04
L	.19	-.41	-.29	.19	-.80	.04	.03	-.02	-.09	.12
S	.09	.17	-.40	-.82	.32	.08	.01	-.00	.03	-.08
USE	.25	-.14	-.82	-.44	-.18	.08	.02	-.02	-.03	.02
INFR	-.01	-.83	.31	-.35	-.21	.18	.08	.01	-.04	.08
POND	-.06	.11	.04	.97	-.08	-.15	-.04	-.00	.00	.05
I	.91	.00	-.17	-.07	-.11	.12	.07	.02	.15	.01
ISD	.91	-.06	-.14	-.03	-.14	.12	.13	.05	.09	-.06
D	-.75	.08	.03	.01	-.13	-.21	.45	.12	.06	.05
TDF	.09	-.04	-.01	-.11	-.15	.06	-.01	-.08	-.95	.06
TPCP	.57	.02	-.08	-.15	-.10	-.02	.69	.03	.21	-.00
API	.47	-.29	-.01	-.02	.05	.06	.68	-.07	-.17	.01
SOLM	-.02	.51	.19	-.44	.16	-.14	.23	.48	-.19	-.01
WDIR	.02	-.31	-.10	.09	-.17	.01	.01	.09	-.07	.91
WEEK	.87	-.02	-.14	.03	-.07	.21	.26	-.08	-.10	.04
AIRT	.90	-.03	-.05	.06	.11	-.14	-.00	-.01	-.13	.12
ATSD	-.36	-.02	-.03	.02	-.00	-.08	-.03	.84	.13	.10
WVEL	-.30	.26	.17	.18	.02	-.76	.18	-.18	.18	.16
WVSD	-.09	.31	-.01	.12	.02	-.82	-.13	.28	-.04	-.14
SOLT	.95	-.08	-.13	-.02	-.08	.09	.10	-.12	-.06	.02
STSD	.27	-.18	-.53	.24	-.37	-.02	.32	.25	-.01	.28
DEGD	.86	-.14	-.13	-.12	-.12	-.08	.07	-.09	-.13	-.09
SWEQ	-.88	-.02	.10	.11	.11	-.04	-.18	.18	-.00	-.06

of each factor is included in the table, showing that the ten rotated factors account for all the variance in the ten components, as expected. The loadings of the factors are the correlations of each variable with each factor,

and the largest correlation for each variable is underlined. As shown, all the variables had a high correlation with only one factor.

From the loadings and the criteria in the literature, the variables were examined for their importance to the rotated factors. As with the choice of the important components, there are no analytical means of determining which variables are important, or which factors could be omitted. The literature stresses the fact that the important variables should be chosen with "one hand over their names", so that the investigator does not force the data to indicate a wrong interpretation. Investigators of multivariate analyses tend to agree that no more than two variables per factor should be retained, and that loadings smaller than about 0.40 can usually be considered to be indicative of an unimportant correlation (Wallis, 1965). This value should only be considered as a guide and not as the accepted criterion for unimportance.

An investigation of Table VII yields several possible interpretations, even though the simple structure is relatively clear because of the single, large loading on each of the variables. Seven of the variables have comparatively small correlations with all the factors, and five of the remaining variables are highly correlated

with factors whose variances are small. The variables which are not highly correlated with any of the factors, GNDL, D, TPCP, API, SOLM, and STSD, do not contribute appreciably to the information contained by the measurements of the original 29 independent variables. Because the rotated factors "reproduce" 93.4 per cent of the original information, omission of the seven variables listed would give approximately the same information contained by measurements of all the variables.

Also, because of the small variances of the last four or five factors, these factors do not reproduce a large portion of the original information. The last three are determined by essentially one variable each, ATSD, TDF, and WDIR, respectively. These three variables are highly correlated with the last three factors, but the factors are not highly important to the total variance, and the variables were therefore deemed to be relatively unimportant.

Two other variables, L and WVSD, were important, respectively, to the fifth and sixth factors. The sixth factor was classified with the last three, and the variable WVSD was also deemed to be unimportant. The relatively large change in variance between the fifth and sixth factors, when compared to the other changes, seems to justify

the single classification of the last five factors, i.e., not nearly as important as the first five.

The variables L and A are the only two variables important to the fifth factor. Because variables which "coexist" within a factor are generally highly correlated, then the exclusion of one yields essentially the same information obtained when both are measured. Because the variable A was considerably more highly correlated with this factor, and because the factor was the least important of the first five, L was also deemed unimportant.

The interpretation given in the preceeding paragraph excludes 12 independent variables. An investigation of the sixth and seventh factors yields the fact that all the variables correlated with these factors have been excluded, and the variances of the factors are essentially equal. Wallis (1965) warns against omitting all the variables important to the important factors, and for this reason, the "hand" was lifted and the variable TPCP in the seventh factor was not omitted in further analyses. Objectivity, and the frequency of appearance of this variable in other investigations prohibited its exclusion here. This essentially means that the first six factors were deemed important, if the sixth and seventh factors in Table VII are reversed. This interpretation left only 18

independent variables for further analysis.

Because a regression equation involving 18 independent variables is probably not practical, two attempts to further reduce the number of variables were made. The first involved the use of only those variables, again with the exception of TPCP, which displayed correlations higher than an arbitrary value of 0.85 with the important factors. The second attempt took advantage of the facts that the "coexistent" variables within a factor are usually highly correlated, and that the factors themselves are uncorrelated. One variable from each of the first six factors, using the reversed notation on factors 6 and 7 in Table VII, was retained to yield six relatively uncorrelated variables. In the first factor, SOLT had the highest correlation; however, for reasons discussed later, the author believed that the variable having the second highest correlation, I, would be a better variable to measure the information contained by the first factor. The variables having the highest correlations with the next five factors were also included, yielding the six most uncorrelated variables. Table VIII shows the three interpretations of the rotated factors. Other interpretations were possible, but the author felt that these would be adequate for the regression analysis to follow. The second and third are meant only

to possibly reduce the number of independent variables, and the first is felt to be the most appropriate interpretation of the rotated factors.

Table VIII. Three Sets of Important Variables from Three Interpretations of Table VII

First Interpretation	Second Interpretation	Third Interpretation
A	A	A
SHP	SHP	SHP
AZ	AZ	
ELEV	ELEV	
GNDS	GNDS	GNDS
FREQ		
S		
USE		
INFR		
POND	POND	POND
I	I	I
ISD	ISD	
TPCP	TPCP	TPCP
WEEK	WEEK	
AIRT	AIRT	
SOLT	SOLT	
DEGD	DEGD	
SWEQ	SWEQ	
18	14	6

If an interpretation of the most important successive variables to the information were desired, Table IX would be the best approach. As shown, the 29 independent variables are all present. The criterion used in formulating this table was the decreasing order of correlations of the variables with the factors from Table VII, with the exception of the first factor, where SOLT was placed third

in line instead of first.

Table IX. Successive Importance of the Variables to the Factors

Factor	Decreasing Order of Importance							
1	I	ISD	SOLT	AIRT	SWEQ	WEEK	DEGD	D
2	AZ	GNDS	ELEV	INFR	SOLM			
3	SHP	USE	FREQ	GNDL	STSD			
4	POND	S						
5	A	L						
6	WVSD	WVEL						
7	TPCP	API						
8	ATSD							
9	TDF							
10	WDIR							

After the interpretations of the initial rotation, the ten original, unrotated factors were rotated three additional times, using only the loadings on the important variables for each of the three interpretations. This was done to refine the rotated loadings on the important variables. In each of the three rotations, the coefficients did not appreciably change, but the new rotated factors were computed to remove any effects of the unimportant variables in each case. All ten of the original factors were rotated to maintain a relatively high percentage of explained variance. If only the first six factors had been rotated, less than 76 per cent of the information would have been retained (see Table VII). The reduction

in the number of variables used in the rotation causes a loss of information, but this is not nearly as much as would have been lost if combined with a further reduction in the number of factors. Table X outlines the retained information for the three rotations. The percentages given are for the original variance in 29 variables, and are therefore small. However, the retained variance of the important variables is large in each case.

Table X . Variance Retained by Rotations of the Original Factors with Reduced Numbers of Variables

Interpretation	Variables	Total Variance of 10 Factors	% of Original Variance
1	18	17.14	88.9
2	14	13.18	87.9
3	6	5.57	86.6

Regression Analyses

Four prediction equations having the form of Equation (3) were obtained. The first used the modified principal component method, and the other three used computed values of the rotated factors from the original data.

Equation (21) was used to find the regression coefficients for the ten important principal components, using the six-place values of the loadings and eigenvalues listed

in Tables V and VI. The computations of Equation (21)

$$F_j = \frac{\sum_{k=1}^{p} l_{kj}\, r_{ky}}{L_j} \tag{21}$$

and the procedure of Equations (17) through (20) were done by a short computer program written for this purpose, giving the coefficients of the measured independent variables. These are given in Table XI at the end of this chapter, along with the results of the other regression analyses. Multiplicative equations for both dependent variables may be obtained from this table and Equation (2)

$$Y = B_0\, X_1^{B1}\, X_2^{B2}\, X_3^{B3} \ldots X_p^{Bp} \tag{2}$$

where B_0 is the antilogarithm of the constant of regression.

An ordinary linear multiple regression analysis was performed for each of the two dependent variables and the ten rotated factors for each of the three interpretations of the initial rotation. A computer program, "One-Pass Multiple Regression", written by the Computing Center Staff for general use, was employed. Input to this program for each analysis was 50 observations of each dependent variable and 50 computed values of ten factors, because the regression was to be performed on the factors and not on

the important variables. The 50 values of the 10 factors for each regression were computed and punched by a program written to comply with Equation (22). The x_{ki} values in this equation are observations of the standardized form of the important, log-transformed variables. The standardized values were computed from Equation (19) for all 50 observations, and only the measurements of the 18, 14, or 6 important variables were used in the computations for the values of the factors.

The results of the regressions were the coefficients for the rotated factors, G_0 and G_j, which had to be reduced to coefficients for the important, log-transformed independent variables. The values of G_0 and G_j were substituted into Equations (23), and a computer program for these computations yielded the values of B_k and B_0 for Equation (2), the desired multiplicative form of the prediction equation. Table XI summarizes the results of all eight regression analyses. The antilogarithms of the constants are given, and any of the prediction equations may be obtained from Equation (2), where the X's are the field measurements of the variables, B_0 is the antilogarithm of the regression constant, and the B_i are the coefficients listed. For example, the equation for the peak discharge rate from the second factor regression is found from Table XI and

Equation (2) to be

$$QMAX = 1.476896 \frac{A^{.263100} \; AZ^{.628922} \; ELEV^{.318249} \; POND^{.091439}}{SHP^{.387924} \; GNDS^{.308629} \; ISD^{.179825}}$$

$$\times \frac{I^{.151434}}{TPCP^{.068278} \; WEEK^{.660111} \; AIRT^{.382531} \; DEGD^{.432310}}$$

$$\times \frac{SOLT^{.302547}}{SWEQ^{.078893}}$$

The above equation was obtained from the fifth column of Table XI, and an equation for the runoff volume may be obtained in a similar fashion from the sixth column. The units for QMAX in the above equation and in the other equations for QMAX are cubic feet per second, and the units for RUNF are inches, providing that the units of the independent variables are the same as the units used in Table II.

Table XI. Coefficients and Constants for the Various Regression Equations

Ind. Var.	Prin.-Comp. Regr. QMAX	Prin.-Comp. Regr. RUNF	Factor Regr. 1 QMAX	Factor Regr. 1 RUNF	Factor Regr. 2 QMAX	Factor Regr. 2 RUNF	Factor Regr. 3 QMAX	Factor Regr. 3 RUNF
A	0.064954	-0.244249	0.598575	-0.053316	0.263100	-0.333236	0.365389	-0.321135
SHP	0.072176	-0.201910	-0.364258	-0.434485	-0.387924	-1.434203	1.953393	2.144183
AZ	0.356270	-0.018289	0.960791	0.109609	0.628922	0.031194		
ELEV	0.384777	0.007589	0.674550	0.024181	0.318249	-0.286868		
GNDS	-0.165029	-0.017250	-0.421549	-0.063942	-0.308629	-0.072376	0.045762	0.946765
GNDL	0.203446	-0.039752						
FREQ	-0.054744	-0.192658	-0.146716	-0.206956				
L	0.032743	-0.312621						
S	-0.118441	-0.140557	-0.513599	-0.231293				
USE	-0.013515	-0.094340	-0.033466	-0.095481				
INFR	-3.363300	-0.140002	-5.579039	0.151752				
POND	0.032398	0.025304	0.071431	0.024675	0.091439	0.071240	0.154250	0.222681
I	0.085569	-0.006457	0.407550	-0.129624	0.151434	-0.133070	-0.339572	-0.360853
ISD	0.043673	-0.015544	0.152213	-0.304537	-0.179825	-0.299678		
D	0.068473	0.204443						
TDF	-0.047350	-0.049478						
TPCP	0.184594	0.324878	-0.063721	0.123028	-0.068278	0.117397	-0.330438	-0.127378
API	0.029977	0.221509						
SOLM	-0.321959	-0.098596						
WDIR	-0.003048	0.275098						
WEEK	0.136325	0.159620	-0.739918	0.271067	-0.660111	0.230633		
AIRT	-0.088573	-0.172147	-0.432860	-0.406769	-0.382531	-0.411448		
ATSD	-0.248782	-0.173896						
WVEL	0.367476	0.519013						
WVSD	-0.374424	-0.902691						
SOLT	0.066545	-0.095602	-0.095502	0.134757	0.302547	0.182688		
STSD	0.120357	0.188501						
DEGD	-0.083005	-0.293114	-0.379431	-0.868418	-0.432310	-0.878633		
SWEQ	-0.013592	-0.007082	-0.012188	-0.060910	-0.078893	-0.063632		
Cst.=	-3.795394	0.201464	-7.323573	-0.532590	0.169350	1.686998	0.337783	-0.418311
Alog=	1.6×10^{-4}	1.590245	4.75×10^{-8}	0.293366	1.476896	48.64045	2.176620	0.381671

Chapter V

DISCUSSION OF RESULTS

The methods employed and the results obtained in the previous chapter indicate that the multivariate procedures provide a statistical means of determining the variables which are more important to runoff than other variables. A discussion of the variables and the prediction equations involving the variables follows.

VARIABLES IMPORTANT TO RUNOFF

The variables which were selected for this study are believed to have been the best possible with the available data. Each possibility was carefully analyzed before the final decisions were made. The possibility of neglecting unknown variables which have considerable importance to the runoff always exists in any hydrologic analysis, because the processes and factors are not yet completely defined (Sharp and Biswas, 1965). The variables were selected according to criteria set forth by Chow (1964). They are grouped into two categories, "climatic" and "physiographic." The climatic factors are sub-classified as those measuring four processes of hydrology: precipitation, interception, evaporation, and transpiration. The physiographic factors have two major sub-sets, basin and channel characteristics.

Practically all of the factors listed by Chow were

present in this study. The precipitation factors of intensity, duration, time distribution, areal distrubution, direction of storm movement, antecedent precipitation, and soil moisture were directly or indirectly measured by the variables I, D, TDF, TPCP, WDIR, API, and SOLM, respectively.

The interception factors were probably not measured as well as the others, but the land-use variable, USE, indirectly measured the vegetation. This variable, combined with the week of the year, and possibly the wind velocity, should give any model "feeling" for the amount of interception which might occur in a storm.

An accurate amount of evaporation is extremely hard to measure, but if the variables which affect evaporation: air temperature, soil temperature, wind velocity, azimuth (North-sloping vs South-sloping), areal extent, pondage, and possibly soil moisture; are measured, then the model has probably been given an estimation of the evaporation which might occur.

Transpiration is relatively unimportant for unvegetated watersheds. The variables USE, SOLM, WEEK, AIRT, WVEL, and SOLT were included to measure vegetation, soil moisture, season, air temperature, wind velocity, and soil temperature, and all of these probably affect the amount

of transpiration which occurs during a runoff event. Also, transpiration is generally negligible during periods of rainfall, and it would be logical to assume that small losses of moisture by this process occurred during the relatively short storm periods used herein. The variables DEGD and SWEQ were included because snow-melt runoff, although a function of many variables, is at least related to the antecedent temperature and the volume of water in the snowpack before an event. Snowmelt runoff events comprised about one-third of the runoff events of this study, and a prediction equation applicable to these was desired.

The physiographic factors of size, shape, slope, orientation, ponds and reservoirs, channel slope, and channel length were measured by the variables A, SHP, GNDS, AZ, ELEV, FREQ, INFR, POND, S, and L, respectively. Some factors suggested by Chow which were not included in this analysis were the frequency of occurrence of precipitation, atmospheric pressure, shape of evaporative surface, solar radiation, humidity, ground water capacity, sediment transport, and channel shape and roughness.

IMPORTANT VARIABLES FROM THE ANALYSES

The variables selected as the most important to each of the factors in Table IX follow the interpretation of the factor rotation (Table VII). The variables I, ISD, and

SOLT in factor 1 were not written according to the factor loadings for two reasons. First, the weather stations were not recording for 14 of the 50 storms, and the treatment for missing data had to be applied to obtain the weather station variables for these storms. The intensities and standard deviation of intensities were measured for all storms, and the author therefore felt justified in placing the soil temperature variable in the third position of importance to this factor. The high loading in this factor for SOLT was therefore probably due to the missing data treatment used for this variable. The average soil temperature for the same time period in other years of measurement may not have been the best choice for the value to assign to missing storms. However, the air temperature was also missing for these storms, and no other alternative was available.

All of the other variable "placings" in Table IX appear to be logical. The position of SOLM in factor 2 might be questioned, not from a hydrologic point of view, but from an error interpretation. The measurements of the soil moisture at any of the three depths were probably the least accurately obtained of all the variables. The soil moisture readings were taken from variations in resistance measured across electrodes embedded in gypsum

blocks, and the resistance readings were converted to water content values by means of calibration charts. The readings did not seem to fluctuate considerably from day to day, and only occasionally did they change during a storm. The calibration was not well established, and some of the soil moisture values used may therefore be questionable. In any event, soil moisture had a small correlation with all the factors, and since no information on the accuracy of the readings was available, the variable was discarded from further analyses. Intuitively, the soil moisture, or at least the rate-change in soil moisture shortly after a storm, would indicate the amount of precipitation that was being lost to the soil. Also, the soil moisture before a storm should indicate the capacity of the soil for receiving the precipitation. This variable should therefore not be excluded from future analyses simply because this study resulted in its being deemed unimportant.

Variables deemed important in each of the three factor-regressions deserve discussion. In the third regression, four of the six variables chosen from the first six factors appear in virtually all regression equations that have been proposed for both peak discharge rate and runoff volume. The watershed area, watershed shape, storm intensity,

and total amount of precipitation are intuitively and, from the results, shown to be mathematically important to run-off. The high importance of the maximum ground slope, GNDS, could be attributed to two possible explanations. Error of measurement may have caused an "apparant" importance, because the end of the blue line on U.S.G.S. maps may not be indicative of the actual main channel head. When this fact is combined with the questionable precision of contours on the maps, a strong possibility for errors in the measurement of this variable exists. An alternative explanation, which justified the importance of this variable, is the supposition that these were "small" watersheds. Because the amount of overland flow rather than channel flow is the criterion (Chow, p. 14-5) for defining a small watershed, then ground slope is more important to peak discharge rates for smaller watersheds. Neither of these reasons for the importance of GNDS in the rotated factors could be selected as the better, and further research is suggested for the solution.

The sixth variable in the third regression equation, pondage, and the main channel slope were the only important variables in the fourth rotated factor. The great difference between the coefficients for these variables forced the use of POND from this factor rather than S.

Had the correlation of S with this factor been larger, the author would have used S instead of POND in the third factor-regression, because the pondage variable was difficult to measure. Pondage is certainly important to runoff volumes and peak discharge rates, but the method of obtaining this variable should create some doubt about the high importance indicated by the analysis.

All three variables measuring the wind velocity and direction were deemed unimportant by the analysis. The measurements of the variables were fairly accurate, and the indication of the rotation is probably correct. Evaporation may not have been significant because the time from the beginning of precipitation to the end of the runoff hydrograph for most of the storms was relatively short.

Duration and time distribution of the storms were other variables deemed unimportant by the analysis. The author feels that this is probably due to the inclusion of snowmelt events in the analysis. Because the last major snowfall before the respective snowmelt-runoff was used, the duration was generally large for these events, and the time distribution was generally zero. Because peak discharge rates from "small" watersheds are highly sensitive to short duration, high-intensity storms (Chow, 1964), the large durations of snowfall storms probably caused the

reduction in the importance of these two variables. If snowmelt events had been excluded, one of these variables might have been found to be more important to runoff.

Both the variables measuring distance were deemed unimportant by the rotation. The meander length of the main channel, L, is usually not important for "small" watersheds, because overland flow rather than channel flow is more important to peak discharge rates. The correlation of 0.80 of this variable with the fifth factor in Table VII was relatively large for rejection, but the importance of factor 5 to the variance makes this correlation less important than an equal correlation in one of the first four factors. The variable is obviously important to the information, but the quest for as few variables as possible for the prediction equation led to its exclusion in the regression. The overland distance, GNDL, was not nearly as important as the main channel distance, which is contrary to the previous definition of "small" watersheds. However, as with the overland slope, this may have been difficult to accurately measure, giving a possible reason for the apparent unimportance of GNDL.

VARIABLE INTERCORRELATIONS

The factors of Table VII indicate that certain names

may be applied to some of the factors. Also, because the factors are uncorrelated, the important variables of one factor are relatively uncorrelated with the variables of another factor. The development also shows that the variables within a factor are usually related. All of these facts may be used to provide information about the intercorrelations of the variables.

The first factor was highly associated with storm variables and snowmelt variables, and might be called a "snowmelt" factor. As shown, the duration of the storm is correlated only with this factor, justifying the earlier statement that the inclusion of snowmelt events affected the duration variable. Factors 6, 7, 8, 9, and 10 could be assigned names such as "wind," "total precipitation," "deviate air temperature," "storm time distribution," and "wind direction," even though the variance of each factor was small.

Factors 2, 3, 4, and 5 are not as easily named as the rest. They are all highly associated with physiographic variables only, indicating that they explain the watershed characteristics, and that the characteristics are unrelated to the storm factors because they are in different factors. This simply means that knowledge of the characteristics of a watershed usually gives no insight into the

storm characteristics.

In the second factor, the soil variables are prevalent. The soil moisture, infiltration rate, and ground slope could very well be related. The first has a definite effect on infiltration rate (Chow, 1964), and the ground slope would also reflect the amount of moisture retained by the soil. The additional presence of azimuth and elevation in this factor led the author to an investigation of the soil types as the elevation changed from watershed to watershed. Although the soil types were not remarkably different, the infiltration rate decreased as the elevation increased, indicating that these variables may be related. The correlation coefficient for the actual measurements of these variables was 0.66, and this was among the largest values for all the correlations. Also, the azimuth variable was highly correlated with both the elevation and infiltration variables. It seems reasonable to state that a North or West-sloping watershed might contain different soils and infiltration rates than a South or East-sloping basin. Prevailing winds, glacier movements, or other factors could easily cause these differences.

Factor 3 indicates that the overland distance, stream frequency, watershed shape, and land use are related.

Because the stream frequency varies approximately as the square of the "drainage density" (Chow, 1964), which is a measure of the closeness of channel segments, and because the overland flow distance is inversely related to drainage density (Chow, 1964), then the correlation of GNDL and FREQ in this factor is justified by the literature. Also, the shape of Montana watersheds may in fact be related to the overland slope and stream frequency, although no statements to this effect could be found in the literature. The presence of USE in this factor seems to have no reasonable explanation.

The fourth factor indicates that the main channel slope and the pondage variables are related. Although the correlation coefficient of these variables in the actual measurement form was small, a coefficient of -0.85 was obtained for the logarithmic forms, indicating a relatively linear plot on logarithmic paper. The negative correlation indicates that more ponds and reservoirs were found on watersheds having milder slopes. This could be attributed to the fact that the "steepest" watershed, Hump Creek, had no visible reservoirs on the maps.

The close relationship of area and main channel length is clearly indicated by the fifth factor, where these are the only two variables highly correlated.

This relationship has been known to exist for several years (Chow, 1964), and the rotation simply adds justification to the relationship.

The remaining factors indicate that the average wind velocity is related to the standard deviation of wind velocities, which is not difficult to accept. The relationship between total precipitation and antecedent precipitation in the seventh factor is interesting. A positive correlation for these variables indicates that a "large" storm was typically preceeded by a "large" amount of precipitation in the 14 days prior to the beginning of the storm. The actual beginning of the runoff-producing storm was sometimes difficult to determine, and this may be the reason for the relationship between these variables. A runoff-producing storm was frequently observed a day or two after another smaller rainfall which apparently produced no runoff. The time of concentration, or the time for rainfall at the farthest reach of the watershed to arrive at the runoff gaging station, assuming a uniform intensity over the watershed, was estimated to be less than one or two days even on the largest watershed. The storms which had no antecedent help usually produced runoff in only a few hours, and this criterion was used in selecting the beginning point for the precipitation

which caused the runoff.

As shown, the methods employed seem to substantiate known relationships of certain variables. Some new relationships for the locale of the watersheds were discovered, and considerable insight into the importance of some of the variables has been obtained. The "simple structure" of the rotated factors in Table VII provided the information discussed, and the agreements with the literature have positive indications of the adequacy of the methods.

REGRESSION EQUATIONS

The coefficients for the variables in the regression equations for total runoff and peak discharge rate were listed in Table XI for the four chosen combinations of variables. The dimensional "imbalance" of the equations is accounted for by the constant for each equation. If units other than those in the derivation are used, then additional conversion constants must be added.

Peak Discharge Rate

In the first equation for peak discharge rate, the coefficients on infiltration rate, both slopes, soil moisture, snow-water equivalent, air temperature, degree

days, and stream frequency were negative, indicating that the peak discharge rate increases as these variables decrease. It would seem that a larger slope should produce larger peak discharge rates, because the velocity of overland and channel flow is directly related to slope. Because the peak discharge rates were measured in the channel, the size, shape, and roughness of the channel are also determinants, and the slope cannot be considered alone. Also, soil moisture and stream frequency are variables which should vary directly with peak discharge rate. Negative coefficients on "obvious" variables such as area, intensity, or total precipitation were not present, and this indicates that "reasonable" coefficients were obtained for most of the variables. The positive coefficient for pondage is questionable, because reservoirs are generally installed to reduce peak discharge rates. If measurements of all the variables in this equation are available, then it should give the best prediction for peak discharge rate because it explains 93.4 per cent of the variation in the original data, regardless of the respective signs of the exponents.

The three factor-regression equations for peak discharge rate each had some intuitively incorrect signs for the exponents. All three indicated that total precipitation is inversely related to peak discharge rates,

and only two indicated a direct relation to intensity. This would mean that short-duration, high-intensity storms produced the greatest discharge rates, confirming the statement that these were "small" watersheds. The negative coefficient for total precipitation seems intuitively incorrect, but the definition of small watersheds allows an inverse relationship of peak discharge rate with duration, and probably with total precipitation because this generally increases with duration.

As in the principal component regression equation, the pondage variable for the last three equations was directly related to peak discharge rate. The reason for this apparent error is unexplained, and deserves further investigation if the signs of the coefficients are accepted as being correct.

The area was found to be directly related to peak discharge rate in all the equations, agreeing with the Rational equation for runoff. However, the ground slope coefficient was positive only in the last equation, and the lack of information of the channel shape and roughness might have been at fault here. Also, this may have been caused by the fact that few runoff events with large peak discharge rates were included for the "steepest" watershed, Hump Creek.

Other coefficients for variables in the equations for peak discharge rate could not be satisfactorily evaluated, because too little was known about their physical influence on discharge. The peculiarities of the research watersheds and measured storms may have resulted in some of the incorrect signs. The equations are not as sensible as expected, but the theory states that because they were obtained from uncorrelated variates, they at least satisfy regression requirements.

Total Runoff

The equations for total runoff also exhibited intuitively incorrect signs for a few coefficients. The most obvious was the inverse relationship of total runoff and watershed area in all the equations. This indicates that larger runoff volumes were produced from smaller watersheds, and may be due to the fact that the largest runoff volume occurred from a storm on Hump Creek. This storm had a particularly long duration with a small peak discharge rate and a large total volume of runoff. The possible influence of one storm on the final equations indicates that the method might be overly sensitive to individual storms, but this statement was not supported by the literature. Further studies are suggested if these methods are to be employed in future regression analyses.

The ground slope was inversely related to runoff volume for the first three equations, and this might be viewed as incorrect, because the milder slopes should retain more precipitation, thereby reducing the total runoff. However, infiltration rates and antecedent soil-moisture conditions must also be considered. The slope cannot be separated from other variables in determining the correct relationship with runoff, and the signs obtained may be correct.

Another possible intuitive error in the equations for runoff is the signs for the coefficients of intensity of precipitation, which are negative for all equations. High intensity storms should reasonably produce large runoff volumes; however, other variables such as duration and the snowmelt variables must also be considered. The low-intensity snow storms with long durations could have predeeded large runoff events, giving a possible explanation for the inverse relationship.

The total precipitation should obviously be directly related to total runoff, and this was found to be true in the first three equations. The negative coefficient in the fourth equation should definitely be questioned for this variable. Large snowmelt-runoff volumes for relatively small amounts of total precipitation could have created this error, but this should have caused negative

signs in all the equations and may not be a reasonable explanation.

In summary, it sould seem that the equations for peak discharge rate and total runoff were theoretically accurate, but the sensibility of the coefficients was not as consistent as the literature had predicted. Most of the incorrect signs could be justified by stating that other variables or watershed and storm peculiarities were responsible, but some coefficients could not reasonably be placed in this category. The suggestions for the limitations of these equations in the next section should provide caution in any predictions.

LIMITATIONS OF RESULTS

Statistical conclusions from four years of information on five watersheds may not be reliable. In general, hydrologists prefer a much longer data period. However, no frequency studies were made in this report, and the author felt that the size-range of storms and floods was well represented by the data. Discharge rates ranging from 10 to 1720 cfs, and runoff volumes between 0.006 and 1.806 inches were present. Because no frequency analyses were performed, and because the mechanics of the variables was the prime objective of this study, sufficient data was felt to be present. The measurements of the variables

were made during each runoff event, and measurements of the mechanics of the variables were therefore taken. Had only a single runoff event been observed, then the rotation method should still have indicated the important variables for the storm, and the regression analysis should have yielded an equation which predicts that particular storm. The use of 50 runoff events simply extends the information about the mechanics of the variables. Recognition of the mechanics from the data is attempted in this report, but is not yet complete.

The regression equations are probably not as reliable as the information of the variables. Long term variables such as annual snow fall, solar radiation, prevailing winds, or watershed "treatments" such as conservation practices, crop rotation, etc., could not be measured in the data period. Prediction with the equations should probably be confined to storms occurring within a time period which is not affected by the long term variables. This period remains to be determined.

With the above limitations, the equations should be suitable for use on ungaged watersheds in the locale of the research watersheds. Measurements or estimations of the independent variables must be made, and some of these may be difficult to obtain, particularly in the

longer equations. The variables in the last equation can be obtained from contoured topographic maps and from estimations of the storm intensity and volume. Design storms will require estimates of the T-year intensities and volumes, which would hopefully give the T-year peak discharge rates and runoff volumes. This is conjecture at this point, and frequency analyses of the data would have to be made before this relationship could be used.

Use of values of the independent variables "outside" the range of each variable is usually discouraged in linear regression equations. The reasoning here is that the equation predicts a line, plane, or hyperplane whose boundaries are the limits of the measured independent variables. For example, a simple linear regression for one variable in terms of another would yield a straight line through the data points on a plot of the two variables. If only a small range of values for the variables was used, the line may be only a segment of a larger line, which may be curvilinear. Prediction outside the range of the variables would yield a point on the extended line, but this line may not represent the total curve. The actual and transformed variable means and standard deviations listed in Appendix H may be used as a guide in determining the range for each variable.

Chapter VI

CONCLUSIONS AND RECOMMENDATIONS

CONCLUSIONS

Two objectives for this investigation were presented in Chapter I. The first was the determination of which of the 29 independent variables were more important to the dependent variables: peak discharge rate and runoff volume from small, central and eastern Montana watersheds. The second objective was the derivation of regression equations for the dependent variables in terms of the important independent variables.

It is believed that the analyses adequately fulfilled the first objective. The variables which contributed most to the peak discharge rates and runoff volumes were in general agreement with the most important variables listed in the literature. The following variables: precipitation intensity, standard deviation of precipitation intensities, soil and air temperature, watershed azimuth, overland slope, watershed shape, reservoir area, and watershed area were among the most successively important independent variables. The variables related to snowmelt runoff events were also important, indicating that some variables were important or unimportant because snowmelt runoff events were included in the investigation.

The principal-component and rotated-factor regression equations for the peak discharge rate and runoff volume from the watersheds were not as consistent as expected. The equations were derived from independent components and factors, and therefore satisfy multiple regression assumptions. The equations for the dependent variables in terms of 29 independent variables exhibited the most intuitively correct relationships among the variables. When variables were discarded for the rotated-factor regression equations, some relationships of dependent and independent variables became intuitively incorrect, especially when only six independent variables were used.

RECOMMENDATIONS FOR FUTURE RESEARCH

Because Wallis (1968) published his paper after the present investigation, some of his suggestions were not included herein, and are presented as recommendations for continued research. The following analyses are suggested to comply with his recommendations: From the rotated factors of Table VII, select only two variables from the first and second factors, and one variable from the third, fourth, fifth, and seventh factors, yielding eight independent variables, I, AIRT, AZ, GNDS, SHP, POND, A, and TPCP, respectively. Perform a principal component analysis of the correlations of only these variables, compute

the principal factors, and rotate the factors. If any of the variables are unimportant to all the rotated factors, repeat the analysis for the important variables. If all the factors are important, compute the principal component regression equation for these variables. This analysis may or may not yield a better prediction equation than any of those included, but the results should prove interesting.

Because the inclusion of snowmelt runoff events in the present investigation caused some of the incorrect signs in the regression equations, the separation of snowmelt and rainfall-produced runoff events is recommended in future analyses. The methods herein are suggested for use in determining important variables to each type of runoff event.

The information about the important variables and related variables could be used in any future study. Although the termination of the measurements of the unimportant variables is not recommended, the author does suggest that concern for precise measurements of many variables may be "relaxed" in favor of better measurements of the important variables. The amount of data used and the use of this single analysis indicate that absolute determinations of totally unimportant variables are not contained herein. This report should serve as a guide for any

future analyses which question the relative importance of certain variables, the possible relationships of certain variables with each other, or the prediction equations for the dependent variables.

SUMMARY

The analysis of peak discharge rates and runoff volumes from five central and eastern Montana watersheds for the important causative factors and the principal component and factor regression equations is contained in this report. Prediction of peak discharge rates and runoff volumes for ungaged watersheds in the locale of the research watersheds is recommended with the developed equations if certain limitations are recognized. The results obtained and the agreements with the literature suggest that multivariate statistical methods are suitable for use in "multi-variable" hydrologic studies for obtaining information about the relative importance of each of the variables, information about the intercorrelations of the variables, and satisfactory regression equations for the runoff variables.

APPENDIX

APPENDIX A

Graphical Derivation of Principal Component Theory

The procedure for finding the l_{kj} in Equation (4) can be viewed graphically to give a physical concept of variance. From Equation (4), the jth component is seen to be linearly related to the standardized variables. If the variables are considered as axes in a p-dimensional space, then any principal component is simply a line through this set of axes. One equation of a line in p dimensions is (Kendall, 1957)

$$\frac{x_1 - m_1}{l_1} = \frac{x_2 - m_2}{l_2} = \ldots = \frac{x_p - m_p}{l_p} \qquad \text{(A1)}$$

where the m_k are intercepts with the x_k axes, and the l_k are the direction cosines. A two-dimensional example of this form is shown in Figure A1, where l_1 and l_2 are the cosines of angles $\emptyset_1$ and $\emptyset_2$. The notation for direction cosines, l, and for the coefficients in Equation (4) is the same because these values are later shown to be the same.

The measurements of the variables, after being standardized, could be plotted as points in the p-dimensional space, just as points P_i are shown in Figure A1. Finding the component which reproduces a maximum amount of the

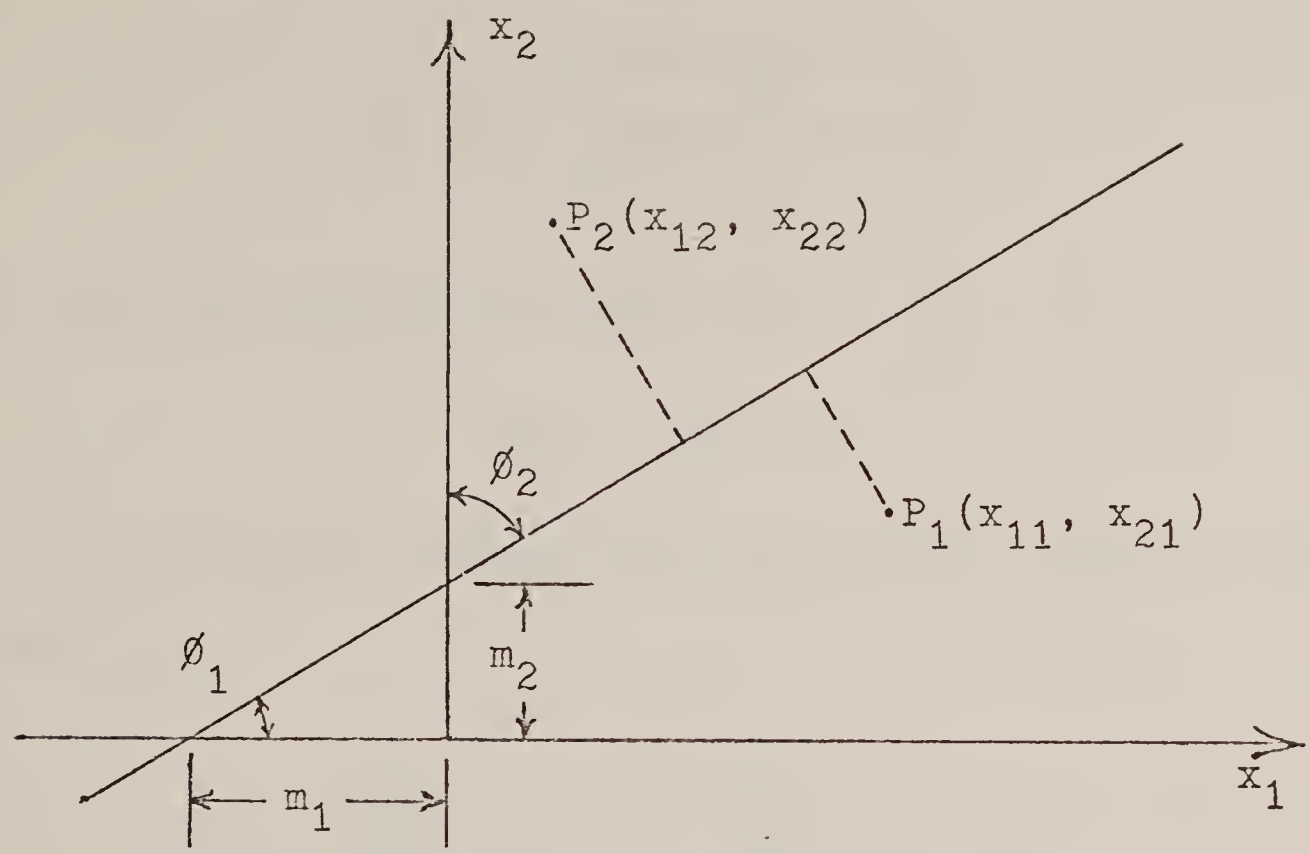

Figure A1. Two-dimensional Principal Component

information is identical to drawing a line through the points in a manner which minimizes the perpendicular distances of the points to the line. The first impulse is to write an equation for the sum of the distances and then minimize this sum. However, some of the distances are negative, and cancellation effects are encountered. To rectify this, the distances can be squared, eliminating the negative signs, and the sum of squared distances can be minimized, which is simply least-squares theory.

Mathematics textbooks give the equation for the distance from point P_i to a line in two dimensions as

$$d_i = \left| \frac{Ax_{1i} + Bx_{2i} - C}{\sqrt{A^2 + B^2}} \right|$$

if the line has the equation

$$Ax_1 + Bx_2 = C \tag{A2}$$

The equation chosen for the line in Figure A1 reduces to

$$\frac{l_2}{l_1} x_1 - x_2 = \frac{l_2}{l_1} m_1 - m_2$$

which has the form of Equation (A2). Substituting for A, B, and C, and squaring d_i gives

$$d_i^2 = \sum_{k=1}^{2} (x_{ki} - m_k)^2 - \left[\sum_{k=1}^{2} l_k (x_{ki} - m_k) \right]^2$$

Since p dimensions instead of two are used in this study, then the squared distance from the $i^{\underline{th}}$ point to the line of Equation (A1) is

$$d_i^2 = S_i = \sum_{k=1}^{p} (x_{ki} - m_k)^2 - \left[\sum_{k=1}^{p} l_k (x_{ki} - m_k) \right]^2$$

and the sum of all squared distances for the n observations is

$$\sum_{i=1}^{n} S_i = nS = \sum_{i=1}^{n} \left(\sum_{k=1}^{p} (x_{ki} - m_k)^2 - \left[\sum_{k=1}^{p} l_k (x_{ki} - m_k) \right]^2 \right) \quad \text{(A3)}$$

which is the form given by Kendall (1957, p. 14). It is desired to find the m_k and l_k which minimize this sum. Upon equating the partial differentials of this equation with respect to the m_k to zero, Kendall shows that the line passes through the origin of the standardized variables, and the m_k are all zero. Before partial differentiation of Equation (A3) with respect to l_k, the condition that

$$\sum_{k=1}^{p} l_k^2 = 1 \quad \text{(A4)}$$

must be included, since the squares of the direction cosines of any line must sum to unity. Although Kendall does not show it, he adds the value

$$nL\left(\sum_{k=1}^{p} l_k^2 - 1\right) = 0$$

to Equation (A3) before differentiating. This is done to assure the investigator that Equation (A4) will apply when the values of l are found. The value L (the eigenvalue) is a constant that may be zero, but it must be

considered unless it is found later to be zero. Dropping m_k in Equation (A3), adding Equation (A5), and differentiating with respect to l_k gives the set of equations

$$\frac{1}{n} \sum_{k=1}^{p} l_k \sum_{i=1}^{n} x_{ki} \, x_{ji} - L \, l_j = 0; \quad j = 1, 2, \ldots, p \qquad \text{(A6)}$$

Since the correlation between two standardized variables is (Harman, 1967, p. 13)

$$r_{jk} = \frac{1}{n} \sum_{i=1}^{n} x_{ji} \, x_{ki} \qquad \text{(A7)}$$

then Equations (A6) can be written as

$$\sum_{k=1}^{p} l_k \, r_{jk} - L \, l_j = 0; \quad j = 1, 2, \ldots, p \qquad \text{(A8)}$$

Noting that the correlation of a variable with itself is unity, Equations (A8) can be expanded to

$$\begin{array}{l}
l_1(1-L) + l_2 \, r_{12} + \ldots + l_p \, r_{1p} = 0 \\
l_1 \, r_{21} + l_2(1-L) + \ldots + l_p \, r_{2p} = 0 \\
\vdots \qquad\qquad\qquad\qquad\qquad\qquad \vdots \\
l_1 \, r_{p1} + l_2 \, r_{p2} + \ldots + l_p(1-L) = 0
\end{array}$$

where the l_k and L are to be determined. These equations can be written in matrix form as

$$|R - LI||l| = 0$$

where $|l|$ is a vector of direction cosines. If all

direction cosines were zero, then all the angles between the desired line and the x_k axes would be right angles. This means that the line would be perpendicular to all the axes. Since this is not likely, the l-vector can be eliminated giving the matrix of Equation (10) in Chapter III.

APPENDIX B

Graphical Derivation of Varimax Rotation Theory

Graphically, the principal components are reference axes for the independent variables. If two components and two independent variables were present, then the components could be represented by the axes V_1 and V_2 in Figure B1. Since the components are linear equations

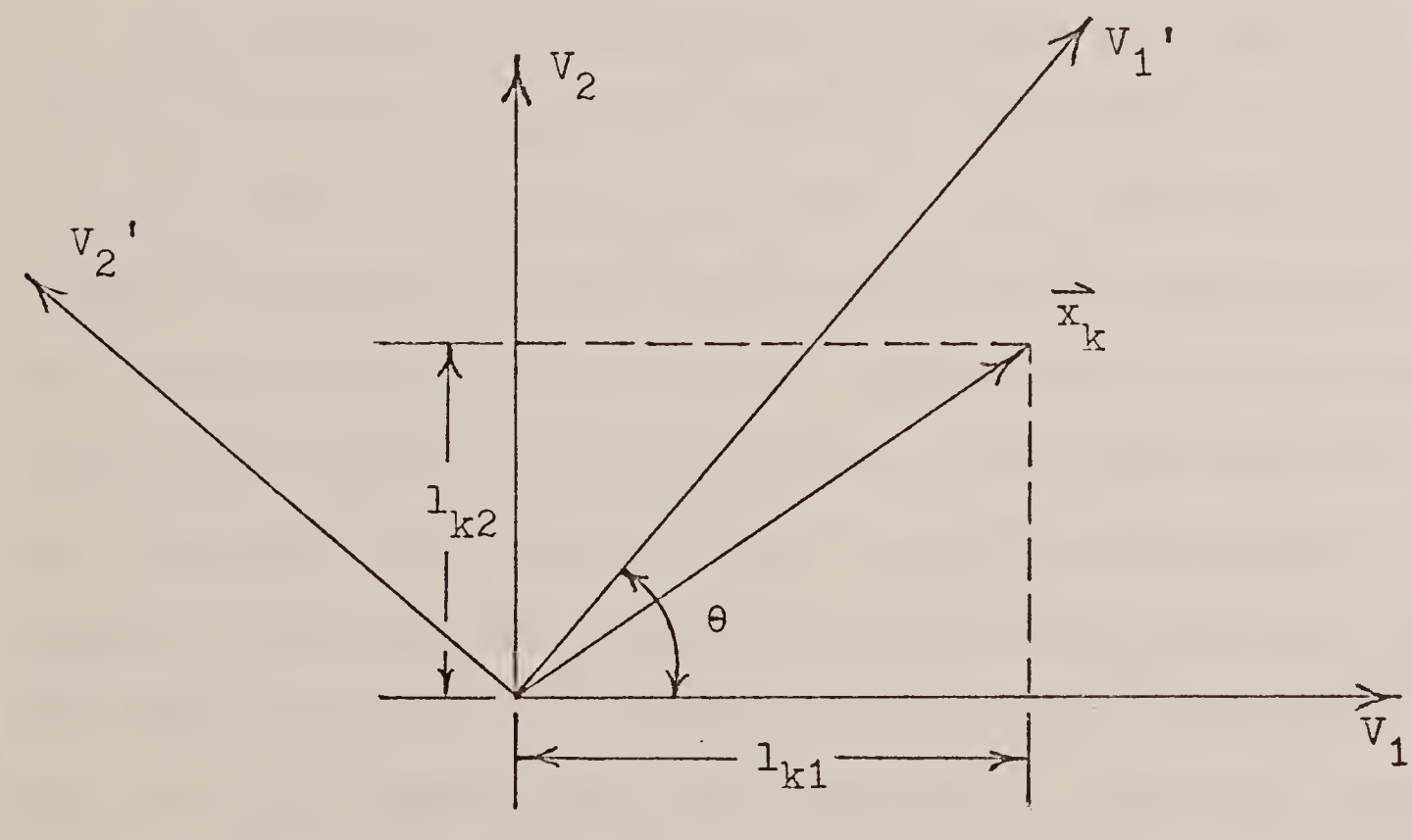

Figure B1. Graphical Representation of Principal Components

involving the standardized independent variables, as shown in Equation (4), then the variables can also be written in terms of the s important components (Kendall,

p. 16). The variable equations are

$$x_{ki} = \sum_{j=1}^{s} l_{kj} V_{ji} \tag{B1}$$

where the terms are the same as for Equation (4). If the V_j are considered to be directional unit vectors, then the l_{kj} are projections of the "variable" vectors, x_k, onto the V_j axes, and Equation (B1) defines the position vector for variable x_k in the reference frame of Figure B1.

The purpose of rotation is to rotate the axes V_1 and V_2 through an angle θ to the new positions, V_1' and V_2', so that the projections onto the new axes are as large as possible. This means that the new components will have either large or small coefficients on the variables, and better interpretations of the importance of the variables will be possible. Coefficients similar to those of the original components (i.e., coefficients that are large for all the variables) will still be present, but if large coefficients are maximized, some small coefficients will be present in the rotated components.

Kaiser obtained his best interpretations from "normal varimax rotation" (1959). In order to use his methods, the components must first be converted to factors by multiplying the coefficients of each component by the component's

eigenvalue, as shown by Equation (12), or

$$A_j = \sqrt{L_j}\ V_j \tag{12}$$

which gives the $j^{\underline{th}}$ factor the equation

$$A_j = L_j \sum_{k=1}^{p} l_{kj}\ x_k$$

if Equation (4) is substituted for V_j. (Baggaley (1964, p. 256) was concerned with the effect of this modification of the components on the interpretation of the variables, and found that nothing was changed. The variance of a factor is the same as the variance of a component (Harman, p. 16), and the contribution of both to the total variance is the same.)

After the factors are found, the variable vectors are given unit length by dividing each coefficient by the length of the vector. The length is simply the square root of the sum of squared coefficients. The new "normalized" variable vectors are therefore defined by

$$Z_k = \sum_{j=1}^{s} a_{kj}\ V_j \tag{B2}$$

for the s components, where

$$a_{kj} = \frac{\sqrt{L_j}\ l_{kj}}{h_k} \tag{B3}$$

and

$$h_k = \sqrt{\sum_{j=1}^{s} L_j (l_{kj})^2}$$

This procedure results in variable vectors having unit lengths and plotted in a reference frame with factors for axes, as shown in Figure B2. The desired

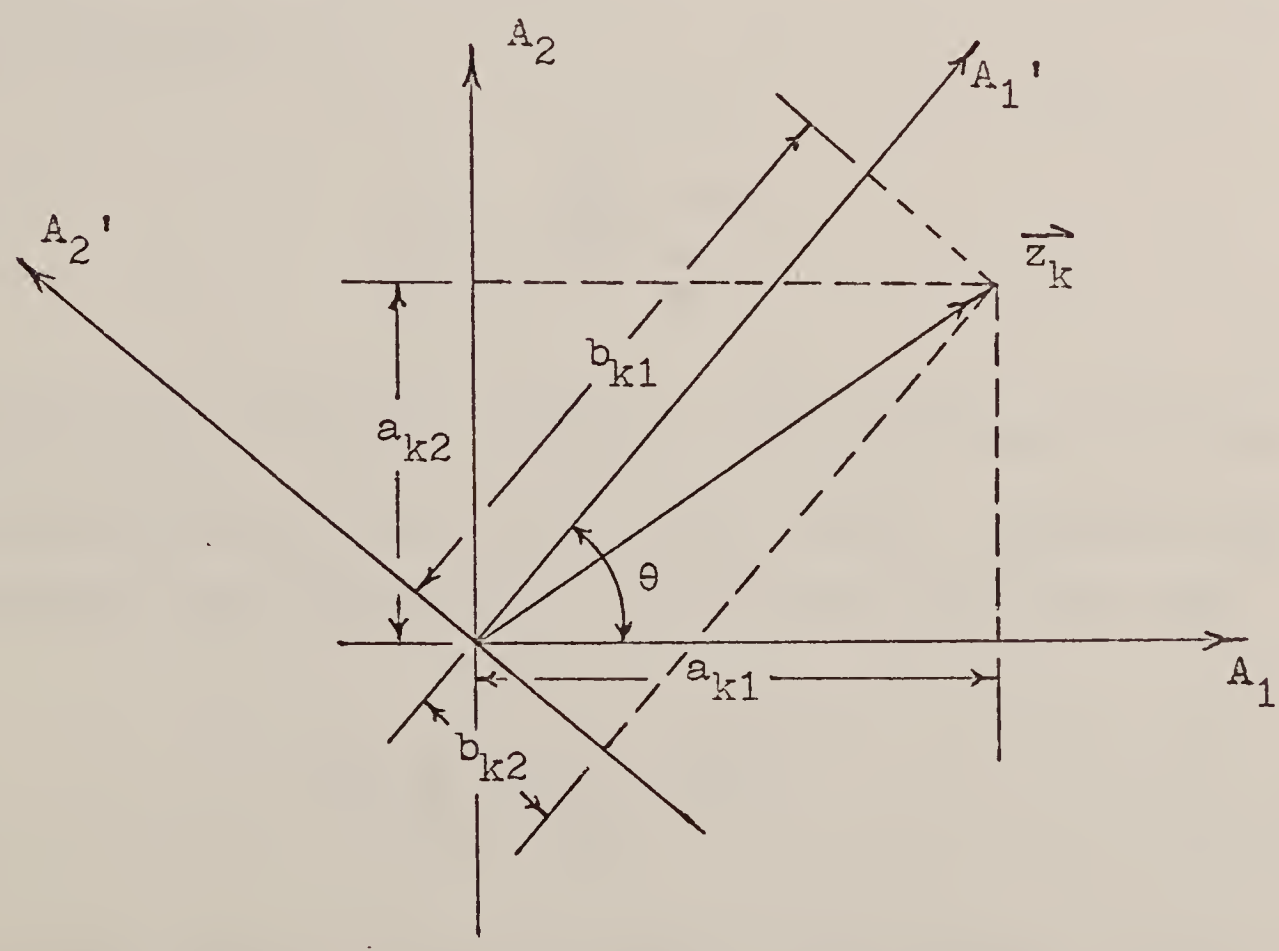

Figure B2. Normalized Variable Vectors Plotted in a Factor Reference Frame

angle of rotation is still θ, and the projections to be maximized are now the b_{kj}, the projections on the rotated factors. As with the principal component derivation in Figure A1, some of the projections may be negative, and squaring eliminates the minus signs. However, the sum of these squared projections is not the variance of a factor, which is to be maximized. Graphically, the variance of a factor is found by subtracting the average squared projection on the $j^{\underline{th}}$ factor from each squared projection, and then squaring this difference to eliminate minus signs. The average squared projection on the $k^{\underline{th}}$ factor is given by

$$\overline{d_k^{\,2}} = \frac{\sum_{k=1}^{p} b_{kj}^{\,2}}{p}$$

if there are p variable vectors. Subtracting this from each squared projection on the $k^{\underline{th}}$ factor, squaring the difference, and summing for all factors gives the sum

$$W = \sum_{j=1}^{s} (b_{kj}^{\,2} - \overline{d_k^{\,2}})^2 \qquad \text{(B4)}$$

This is the sum of all the squared differences for the p variables and s factors, and it represents the sum of

squared projections on the rotated factors. If this sum is divided by p, then an "average" squared projection on the factors is approximated. Kaiser defines the variance of a factor as this "average", and his equation to be maximized is given (1959) by Equation (15) in Chapter III. Equation (15) is obtained by dividing Equation (B4) by p, and expanding the squared term.

After the b_{kj} which maximize Equation (15) are found, the coefficients of each final variable vector are multiplied by the lengths of the original vectors to place them back in the same relative perspective before they were normalized, rather than allowing each to have unit length (Kaiser, 1959). The rotated factors therefore have the final equations

$$A_{ji}' = \sum_{k=1}^{p} c_{kj} x_{ki} \tag{B5}$$

where

$$c_{kj} = h_k b_{kj}$$

and the c_{kj} provide the information for interpretation of the importance of the variables to the rotated factors, and therefore to the information present in the data. The c_{kj} values of Equation (22) are the same as these c_{kj}.

APPENDIX C

Descriptions of Independent Variables

The variable numbers, names, and definitions listed below describe the 29 independent variables of Table II in Chapter IV. Each definition gives the units of the variable in parantheses, the methods of obtaining each variable, and the source of the data for each measurement of each variable.

Physiographic Variables

1. A : Area of watershed (sq mi) from SCS soil map (scale: 4 in. = 1 mi).

2. SHP : Shape of watershed (dimensionless) from the ratio of the length of the longest line that could be passed through the center-of-area of the watershed to the length of a perpendicular through this point. Center-of-area found from the balance point of a cut-out of the watershed from SCS soil map.

3. AZ : Azimuth of main channel (deg) measured clockwise from true North when referred to the water-stage recorder. Mean azimuth of straight-line sections of main channel computed from

$$AZ = \frac{\sum (AZ_i \; L_i)}{\sum L_i}$$

where

AZ_i = azimuth of each straight-line

segment, starting at gaging station and ending at closest point on basin divide to end of main channel on SCS soil map,

and,

L_i = length of each straight line segment.

4. ElEV: Elevation of watershed (ft) computed from an average of the elevations at 0.2 and 0.8 of the main channel straight-line-segment length from the gaging station to the nearest point of basin divide from head of main channel. Elevations from contoured topographic map and distances from chartometer of SCS soil map.

5. GNDS: Maximum ground slope (ft/ft) computed by dividing the distance from the head of the blue, main-channel line on USGS Quadrangle map to the nearest point on the basin divide, by the change in elevation between these points.

6. GNDL: Overland distance (mi) of flow obtained from distance from head of blue, main-channel line on USGS Quadrangle map to nearest point on basin divide.

7. FREQ: Stream frequency (1/sq mi), or number of stream segments per unit area. Obtained by dividing the total number of stream segments of all orders by the area of the watershed. Orders of stream segments determined by numbering heads of tributaties 1st order, and having two 1st order streams joining to a 2nd order, etc. Count taken from SCS soil map.

8. S : Main channel slope (ft/ft) computed as the slope of a straight line drawn through the main channel straight-line-segment profile. The straight line is fitted to the profile in such a way as to give the same area under the line as the total area under the profile, using equal base lengths. Profile from contoured Quadrangle maps (for Duck and E. F. Duck Creek watersheds, profile from transit-stadia surveys).

9. L : Main channel meander length (mi) from chartometer of SCS soil map main channel (largest order stream), including distance from end of blue line to nearest point on basin divide.

10. USE : Land use ratio (dimensionless) obtained from a ratio of the vegetated (including forests for Hump Creek) area to the bare-cultivated (summer fallow) area. Relative areas were determined subjectively from each section in the aerial photographs.

11. INFR: Infiltration rate (in./hr) of watershed, weighted for rates of different soil types. Computed from,

$$INFR = \frac{\sum(I_i \ A_i)}{\sum A_i}$$

where,

I_i = average measured infiltration rate on each soil type of watershed. If not measured, soil was classified by SCS method (type A, B, C, or D), and the average for all watersheds was applied to any soil types which were not tested,

and,

A_i = area of each soil type on watershed from planimeter of SCS soil map.

12. POND: Distance-weighted per cent of total area in ponds and reservoirs (%) computed from,

$$POND = \frac{\sum(L_i \ AP_i)}{A \sum L_i} (100)$$

where,

L_i = straight-line-segment distance along main channel from gaging station to each pond on SCS soil map,

AP_i = full surface area of each pond or reservoir from planimeter of SCS soil map,

and,

A = area of watershed.

Storm Variables

13. I : Intensity of precipitation (in./hr) computed as mean of all hourly intensities from all recording precipitation stations from beginning of storm on the first recording station to hour of last runoff from hydrograph. Beginning of storm determined from precipitation hyetographs. (Last major snowfall used for snowmelt runoff events.)

14. ISD : Standard deviation of precipitation intensities used in variable 13, computed from,

$$(ISD)^2 = \frac{1}{n-1}\left[\sum I_i^{\,2} - \frac{(\sum I_i)^2}{n}\right]$$

where,

n = total number of hourly intensities recorded for all recording precipitation stations.

15. D : Duration of storm (hr) computed as an average time to the center of area of rainfall hyetographs for all recording precipitation stations, computed from,

$$D = \frac{\sum(DCA_i\ A_i)}{\sum A_i}$$

and,

$$DCA_i = \frac{\sum(I_i \; T_i)}{\sum I_i}$$

where,

DCA_i = time in hours from beginning of each rainfall hyetograph to the center of area of each hyetograph, where hyetograph ends with last precipitation before hour of end of runoff,

A_i = Thiessen area of each recording precipitation station from planimeter of SCS soil map,

T_i = time in hours from beginning of hyetograph to I_i, where,

I_i = hourly precipitation at the respective recording station.

16. TDF : Time distribution factor (hr) of precipitation over watershed, computed as,

$$TDF = D1 - D$$

where,

$$D1 = \frac{\sum(DCA_j \; A_j)}{\sum A_j}$$

and,

DCA_j = time in hours from beginning of first appearing rainfall hyetograph to the center of area of each hyetograph,

and,

A_j = Thiessen area of each recording precipitation station from planimeter of SCS soil map.

17. TPCP: Total precipitation (in.) computed as an average Thiessen-weighted precipitation from all precipitation stations, or,

$$TPCP = \frac{\sum(A_i \; P_i)}{\sum A_i}$$

where,

P_i = total storm precipitation at each recording and non-recording precipitation station from beginning of storm to end of runoff,

and,

A_i = Thiessen area of each station from planimeter of SCS soil map.

18. API : Fourteen-day antecedent precipitation index (in.) computed from,

$$API_i = k \; (API_{i-1}) + (P_i)$$

where,

API_i = antecedent precipitation index for the ith day after the fourteenth day before the first day of the storm,

k = 0.78, a reduction constant to indicate the evaporation-transpiration losses from day to day. (Usually ranges from 0.80 to 0.95 (Chow, p. 25-102), and small value used here for relatively dry watersheds.),

P_i = Thiessen-weighted average daily precipitation for each station,

and,

API_{14} = API for first day of storm.

19. SOLM: Average water content (%) of soil for all stations for 3, 9, and 18-inch depths, at beginning of first hyetograph.

20. WDIR: Predominant wind direction during storm period (dimensionless), obtained from weather station data, and given the numbers 1, 2, 3, ..., 8 to represent a wind from the N, NW, W, SW, S, SE, E, NE, respectively.

21. WEEK: Week of the year (dimensionless) in which the date of the peak discharge rate fell. Weeks were numbered from the first week in Jan., regardless of the number of days in this week.

22. AIRT: Average Thiessen-weighted air temperature (deg F), 4 ft above ground, during storm period. Given by,

$$AIRT = \frac{\sum(T_i \; A_i)}{\sum A_i}$$

where,

T_i = mean hourly temperature at a station during the time from beginning of first precipitation to end of runoff hydrograph,

and,

A_i = Thiessen area of each weather station on the watershed.

23. ATSD: Average Thiessen-weighted standard deviation of air temperatures (deg F), 4 ft above ground, during the storm period. Given by,

$$ST_i^{\,2} = \frac{1}{n-1}\left[\sum T_i^{\,2} - \frac{(\sum T_i)^2}{n}\right]$$

and,

$$ATSD = \frac{\sum(ST_i \ A_i)}{\sum A_i}$$

where,

ST_i = standard deviation of hourly air temperatures at a weather station during the storm period,

n = number of hourly air temperatures considered,

T_i = hourly air temperature at 4 ft above ground at a station,

and,

A_i = Thiessen area of each weather station on the watershed.

24. WVEL: Mean wind velocity (mph) during the storm period.

25. WVSD: Standard deviation (mph) of hourly wind velocities during the storm period.

26. SOLT: Average Thiessen-weighted soil temperature (deg F), 3 in. below surface, during storm period. Computed exactly as air temperature mean, only using four readings per day at 6-hr intervals, beginning at 3:00 a.m.

27. STSD: Average Thiessen-weighted standard deviation of soil temperatures (deg F), 3 in. below surface, during storm period. Computed in the same manner as the air temperature standard deviation.

28. DEGD: Average Thiessen-weighted Degree-Days (deg F) for a 14-day period prior to the beginning of runoff. Degree-Days for each day was maximum daily temperature, and 14-day average was,

$$DEGD = \frac{\sum (DD_i \ A_i)}{14 \sum A_i}$$

where,

DD_i = total Degree-Days for a weather station for a 14-day period before the first day of runoff,

and,

A_i = Thiessen area of each weather station on the watershed.

29. SWEQ: Snow-water equivalent (in.), or the volume of water in the snowpack on the watershed on the date of the beginning of runoff. Computed from Thiessen-weighted volume from each snow course. Loss or gain in volume from date of snow survey to date of runoff was computed as the difference between accumulated precipitation and accumulated runoff. This was added to the Thiessen-weighted volume on the watershed at the time of the last snow survey, giving the volume on the date of beginning of runoff. Thiessen areas of snow courses were used to give the volume on the date of the survey, or,

$$V = \frac{\sum (VC_i \ A_i)}{\sum A_i}$$

where,

VC_i = snow-water equivalent at a snow course,

and,

A_i = Thiessen area of each snow course on the watershed.

APPENDIX D

```
C
C           THIS PROGRAM COMPUTES VALUES FOR VARIABLES 13, 14,
C     15, 16, AND 17 USING RECORDING AND NON-RECORDING PRECIP
C     DATA IN USWB FORM.
C           INPUT FOR A WATERSHED CONSISTS OF
C     CARD 1    NO. OF RECORDING PRECIP STAS., NO. OF RUNOFF
C               EVENTS, NO. OF NON-RECORDING PRECIP STAS., AREA
C               OF WSHD.
C     CARD 2    THIESSEN AREAS OF RECORDING PRECIP STAS.
C     CARD 3    THIESSEN AREAS OF ALL PRECIP STAS.-SUM TO AREA.
C     CARD 4    LAST FOUR DIGITS OF STA NOS. OF REC. AND NON-
C               REC. PRECIP STAS.
C     CARD 5    DATE OF PEAK DISCHARGE, DATE FIRST RECORDING
C               PRECIP, HOUR FIRST REC. PRECIP, DATE FIRST REC.
C               PRECIP EACH STATION, HOUR FIRST REC. PRECIP
C               EACH STA., DATE FIRST NON-REC. PRECIP EACH STA,
C               DATE END OF RUNOFF, AND HOUR END OF RUNOFF, ALL
C               FOR FIRST EVENT.
C     CARD 6    SAME DATA ON CARD 5 FOR SECOND EVENT, ETC.
C     CARD 7    RECORDING PRECIP DATA IN ORDER OF STAS.-CARD 2.
C     CARD 8    NON-REC.  PRECIP DATA IN ORDER OF STAS.-CARD 2.
C     CARD 9    BLANK CARD TERMINATES PROGRAM.
    1 FORMAT (1X,4A4,5X,I2,5X,I2,5X,I2,5X,F8.2)
    2 FORMAT (1X,7F8.2)
    3 FORMAT (1X,I4)
    4 FORMAT (1X,3I2,5X,3I2,5X,I2,5X,3I2,5X,I2,5X,3I2,5X,3I2,
     15X,I2)
    5 FORMAT (2X,I4,3I2,1X,12A3 )
   60FORMAT (1X, 35H CARDS OUT OF ORDER FOR STATION NO.,I5,
     113H BEGINNING ON, 2(I2,1H/*,I2/19H PROGRAM TERMINATED )
    7 FORMAT (2X,I4,3I2,10X,A4)
  200 FORMAT ( 1H1)
   80FORMAT (10X, 37H VARIABLES 13, 14, 15, 18, AND 19 FOR,
     14A4,10H WATERSHED //15X,5H PEAK,5X,5H MEAN,5X,
     212H VARIANCE OF,5X,11H STD DEV OF,5X,9H THIESSEN,5X,
     39H THIESSEN,5X,13H THIESSEN AVG/15X,5H DATE,3X,
     410H INTENSITY,2X,12H INTENSITIES,4X,12H INTENSITIES,
     5 5X,9H DURATION, 4X,11H TIME DISTR, 5X,11H TOT PRECIP )
   90FORMAT ( 25X, 5H (IN),  7X, 8H (SQ IN),10X,5H (IN), 8X,
     18H (HOURS),6X,8H (HOURS),10X,5H (IN) //)
  100FORMAT(13X,2(I2,1H/),I2,5X,
     132H MISSING DATA FOR STATION NO 240,I3,3H ON,1X,
     22(I2,1H/),I2,2X,27H COMPUTE VARIABLES BY HAND.)
  110FORMAT(13X,2(I2,1H/),I2,5X,
     138H ACCUMULATED PRECIP FOR STATION NO 240,I3,3H ON,1X,
     22(I2,1H/),I2,2X,27H COMPUTE VARIABLES BY HAND.)
```

```
120FORMAT (13X,2(I2,1H/),I2, 4X,F6.4, 7X,F7.5,10X,F7.4, 8X,
  1F6.1,8X,F6.1,9X,F8.4)
13 FORMAT (2X,I4)
  0DIMENSION THA(4),THAN(7),MSTANO(7),NWSHD(4),NYRP(15),
  1NMOP(15),NDAP(15),NYRB(15),NMOB(15),NDAB(15),NHR(15),
  2NYRBB(15,7),NMOBB(15,7),NDABB(15,7),NHRR(15,7),
  3NYRBN(15,7),NMOBN(15,7),NDABN(15,7),NYRE(15),NMOE(15),
  4NDAE(15),NHRE(15),NNDAB(15),NNDABN(15,7),NNDAE(15),
  5XN(15),XSUM(15),XSUMSQ(15),SUMDA(15),SUMA(15)
  0DIMENSION SUMAP(15),SSUMXT(15,7),SUMXT(15,7),SUMPTA(15),
  1SUMX(15,7),SSUMX(15,7),L(24),KKK(15),SUMXSQ(15),VAR(15),
  2STDEV(15),XBAR(15),DCA(15,7),D(15),PCA(15,7),TL(15),
  3PT(15)
   IIK = 0
14 CONTINUE
   READ 1, (NWSHD(J),J=1,4),II,NS,JJ,A
C     READ STORM CONTROL CARDS FOR A WATERSHED AND CONVERT
C     DATES TO DAY OF THE YEAR.
   IF(A)159,159,15
15 READ 2,(THA(J),J=1,II)
   LL=II+JJ
   READ 2,(THAN(J),J=1,LL    )
   DO 68 KK=1,LL
68 READ 3,MSTANO(KK)
   DO168 K=1,NS
   DO  67 KK=1,LL
  0READ 4,NYRP(K),NMOP(K),NDAP(K),NYRB(K),NMOB(K),NDAB(K),
  1NHR(K),NYRBB(K,KK),NMOBB(K,KK),NDABB(K,KK),NHRR(K,KK),
  2NYRBN(K,KK),NMOBN(K,KK),NDABN(K,KK),NYRE(K),NMOE(K),
  3NDAE(K),NHRE(K)
   NAA=NMOB(K)
   IF(NAA) 172,172,173
173 GO TO (16,17,18,19,20,21,22,23,24,25,26,27),NAA
16 NNDAB(K)=NDAB(K)
   GO TO 28
172 NNDAB(K)=0
   GO TO 28
17 NNDAB(K)=NDAB(K)+31
   GO TO 28
18 NNDAB(K)=NDAB(K)+59
   GO TO 28
19 NNDAB(K)=NDAB(K)+90
   GO TO 28
20 NNDAB(K)=NDAB(K)+120
   GO TO 28
21 NNDAB(K)=NDAB(K)+151
   GO TO 28
```

```
 22 NNDAB(K)=NDAB(K)+181
    GO TO 28
 23 NNDAB(K)=NDAB(K)+212
    GO TO 28
 24 NNDAB(K)=NDAB(K)+243
    GO TO 28
 25 NNDAB(K)=NDAB(K)+273
    GO TO 28
 26 NNDAB(K)=NDAB(K)+304
    GO TO 28
 27 NNDAB(K)=NDAB(K)+334
 28 NBB      =NYRB(K)-62
    IF(NBB)32,32,174
174 GO TO (32,29,32,32,32,29),NBB
 29 IF(NNDAB(K)-60)32,30,30
 30 IF(NMOB(K)-2)32,32,31
 31 NNDAB(K)=NNDAB(K)+1
 32 NCC=NMOBN(K,KK)
    IF(NCC)170,170,171
171 GO TO (33,34,35,36,37,38,39,40,41,42,43,44),NCC
 33 NNDABN(K,KK)=NDABN(K,KK)
    GO TO 45
170 NNDABN(K,KK)=0
    GO TO 45
 34 NNDABN(K,KK)=NDABN(K,KK)+31
    GO TO 45
 35 NNDABN(K,KK)=NDABN(K,KK)+59
    GO TO 45
 36 NNDABN(K,KK)=NDABN(K,KK)+90
    GO TO 45
 37 NNDABN(K,KK)=NDABN(K,KK)+120
    GO TO 45
 38 NNDABN(K,KK)=NDABN(K,KK)+151
    GO TO 45
 39 NNDABN(K,KK)=NDABN(K,KK)+181
    GO TO 45
 40 NNDABN(K,KK)=NDABN(K,KK)+212
    GO TO 45
 41 NNDABN(K,KK)=NDABN(K,KK)+243
    GO TO 45
 42 NNDABN(K,KK)=NDABN(K,KK)+273
    GO TO 45
 43 NNDABN(K,KK)=NDABN(K,KK)+304
    GO TO 45
 44 NNDABN(K,KK)=NDABN(K,KK)+334
 45 NDD          =NYRBN(K,KK)-62
    IF(NDD)49,49,175
```

```
175 GO TO (49,46,49,49,49,46),NDD
 46 IF (NNDABN(K,KK)-60)49,47,47
 47 IF (NMOBN(K,KK)-2)49,49,48
 48 NNDABN(K,KK)=NNDABN(K,KK)+1
 49 NEE=NMOE(K)
    IF(NEE)176,176,177
177 GO TO (50,51,52,53,54,55,56,57,58,59,60,61),NEE
 50 NNDAE(K)=NDAE(K)
    GO TO 62
176 NNDAE(K)=0
    GO TO 62
 51 NNDAE(K)=NDAE(K)+31
    GO TO 62
 52 NNDAE(K)=NDAE(K)+59
    GO TO 62
 53 NNDAE(K)=NDAE(K)+90
    GO TO 62
 54 NNDAE(K)=NDAE(K)+120
    GO TO 62
 55 NNDAE(K)=NDAE(K)+151
    GO TO 62
 56 NNDAE(K)=NDAE(K)+181
    GO TO 62
 57 NNDAE(K)=NDAE(K)+212
    GO TO 62
 58 NNDAE(K)=NDAE(K)+243
    GO TO 62
 59 NNDAE(K)=NDAE(K)+273
    GO TO 62
 60 NNDAE(K)=NDAE(K)+304
    GO TO 62
 61 NNDAE(K)=NDAE(K)+334
 62 NGG     =NYRE(K)-62
    IF(NGG)66,66,178
178 GO TO (66,63,66,66,66,63),NGG
 63 IF(NNDAE(K)-60)66,64,64
 64 IF(NMOE(K)-2)66,66,65
 65 NNDAE(K)=NNDAE(K)+1
 66 XN(K)=0.
    XSUM(K)=0.
    XSUMSQ(K)=0.
    SUMDA(K)=0.
    SUMA(K)=0.
    SUMAP(K)=0.
    KKK(K)=0
    SSUMXT(K,KK)=0.
    SUMXT(K,KK)=0.
```

```
      SUMPTA(K)=0.
      SUMX(K,KK)=0.
   67 SSUMX(K,KK)=0.
  168 CONTINUE
      K=1
      I=1
   69 XHR=0.
      NC=0
      ZHR=0.
      NDAY=0
C     READ AND STORE PERTINENT RECORDING PRECIP DATA.
   70 NDDAY=NDAY
      READ 5, NSTANO,NYR,NMON,NDATE,(L(J),J=1,12)
      READ 5, JSTANO,KYR,KMON,KDATE,(L(J),J=13,24)
      IF(KMON-NMON)73,71,73
   71 IF(KDATE-NDATE)73,72,73
   72 IF(KYR-NYR)73,74,73
   73 PRINT 6,NSTANO,NMON,NDATE,NYR
      GO TO 159
   74 NR=0
      IF(NSTANO-MSTANO(I))70,75,70
   75 GO TO (76,77,78,79,80,81,82,83,84,85,86,87),NMON
   76 NDAY=NDATE
      GO TO 88
   77 NDAY=NDATE+31
      GO TO 88
   78 NDAY=NDATE+59
      GO TO 88
   79 NDAY=NDATE+90
      GO TO 88
   80 NDAY=NDATE+120
      GO TO 88
   81 NDAY=NDATE+151
      GO TO 88
   82 NDAY=NDATE+181
      GO TO 88
   83 NDAY=NDATE+212
      GO TO 88
   84 NDAY=NDATE+243
      GO TO 88
   85 NDAY=NDATE+273
      GO TO 88
   86 NDAY=NDATE+304
      GO TO 88
   87 NDAY=NDATE+334
   88 NNYR=NYR-62
      GO TO (92,89,92,92,92,89),NNYR
```

```
 89 IF(NDAY-60)92,90,90
 90 IF(NMON-2)92,92,91
 91 NDAY=NDAY+1
 92 IF(NR)135,93,135
 93 IF(NDAY-NNDAB(K))94,160,94
160 IF(NYR-NYRB(K))70,103,70
 94 IF(NC-1)95,101,95
 95 IF(NDAY-NNDAB(K))70,70,161
161 IF(NYR-NYRB(K))70,96,70
 96 IF(NDAY-NNDAE(K))98,97,98
 97 XNHRE=NHRE(K)
    SUB=XNHRE-1.
    GO TO 100
 98 IF(NDAY-NNDAE(K))99,99,107
 99 SUB=0.0
100 XDAYS=NDAY-NNDAB(K)
    NC=1
    ZHR=ZHR+24.*XDAYS-SUB
    GO TO 103
101 IF(NDAY-(NDDAY+1))102,103,102
102 XNDAY=NDAY-NDDAY-1
    ZHR=ZHR+24.*XNDAY
    XHR=XHR+24.*XNDAY
103 NC=1
    IF(NDAY-NNDAE(K))105,104,105
104 JK=NHRE(K)
    GO TO 114
105 IF(NDAY-NNDAE(K))106,106,107
106 JK=24
    GO TO 114
107 IF(K-NS)108,111,108
108 IF(NDAY-NNDAB(K+1))109,110,109
109 IF(NDATE-NDABB(K+1,I))128,110,128
110 K=K+1
    ZHR=0.0
    XHR=0.0
    GO TO 103
111 IF(I-II)112,128,112
112 IF(NDATE-NDABB(1,I+1))128,113,128
113 K=1
    I=I+1
    GO TO 103
114 DO 127J=1,JK
    IF(NDAY-NNDAB(K))115,118,115
115 IF(NDATE-NDABB(K,I))120,116,120
116 IF(J-NHRR(K,I))117,120,120
117 ZHR=ZHR+1.0
```

```
      GO TO 127
  118 IF(J-NHR(K))127,119,119
  119 IF(J-NHRR(K,I))121,120,120
  120 XHR=XHR+1.0
  121 ZHR=ZHR+1.0
      K1=L(J)/10000
      NKK=K1*10000
      J1=(L(J)-NKK)/100
      J3=K1/100
      KKKK=J3*100
      J2=K1-KKKK
      IF(J1)122,122,126
  122 IF(J2)124,124,123
  123 ZZ=1.
      KKK(K)=K
      NOO=NDATE
      NOOO=NYR
      NOON=NMON
      NONN=NSTANO
      GO TO 128
  124 IF(J3)125,125,127
  125 ZZ=0.
      KKK(K)=K
      NOO=NDATE
      NOOO=NYR
      NOON=NMON
      NONN=NSTANO
      GO TO 128
C     COMPUTE SUMS AND SUMS OF SQUARES FOR RECORDING DATA.
  126 N=(J1-70)+(J2-70)*10+(J3-70)*100
      X=N
      XXN=X/100.
      SUMXT(K,I)=SUMXT(K,I)+XXN*XHR
      SSUMXT(K,I)=SSUMXT(K,I)+XXN*ZHR
      SSUMX(K,I)=SSUMX(K,I)+XXN
      XN(K)=XN(K)+1.0
      XSUM(K)=XSUM(K)+XXN
      XSUMSQ(K)=XSUMSQ(K)+XXN**2
      SUMX(K,I)=SUMX(K,I)+XXN
  127 CONTINUE
      GO TO 70
  128 IF(K-NS)129,130,129
  129 K=K+1
      GO TO 131
  130 I=I+1
      K=1
  131 IF(I-II-1) 69,132,132
```

```
  132 K=1
      I=1
C     READ AND STORE ALL PERTINENT NON-RECORDING DATA FOR ALL
C     STORMS,
  133 READ 7,NSTANO,NYR,NMON,NDATE,NDSUM
      MM=II+I
      IF(NSTANO-MSTANO(MM  ))133,134,133
  134 KK=II+I
      NR=1
      GO TO 75
  135 K1=NDSUM/100
      NKK=K1*100
      J1=NDSUM-NKK
      K2=K1/100
      NKK1=K2*100
      J2=K1-NKK1
      J4=K2/100
      NKK2=J4*100
      J3=K2-NKK2
      IF(J1-20)138,136,137
  136 XXN=0.0
      GO TO 140
  137 N=(J1-70)+(J2-70)*10+(J3-70)*100+(J4-70)*1000
      X=N
      XXN=X/100.
      GO TO 140
  139 KKK(K)=K
      ZZ=0.0
      NOO=NDATE
      NOOO=NYR
      NOON=NMON
      NONN=NSTANO
      GO TO 144
  138 IF(J4-20)140,136,137
  140 IF(NDAY-NNDABN(K,KK))133,162,162
  162 IF(NYR-NYRBN(K,KK))133,141,133
  141 IF(NDAY-NNDAE(K))142,143,143
  142 IF(J1-20)899,900,900
  899 IF(J4-20)139,900,900
  900 SUMX(K,KK)=SUMX(K,KK)+XXN
      GO TO 133
  143 IF(NDAY-NNDAE(K))144,142,144
  144 IF(K-NS)145,146,145
  145 K=K+1
      GO TO 147
  146 K=1
      I=I+1
```

```
C     COMPUTE VALUES OF VARIABLES FOR EACH STORM.
  147 IF(I-JJ-1)133,148,148
  148 DO 150 K=1,NS
      DO 149 I=1,II
      SUMXSQ(K)=XSUM(K)**2
      IF(XN(K))149,149,404
  404 VAR(K)=(XSUMSQ(K)-SUMXSQ(K)/XN(K))/(XN(K)-1.0)
      STDEV(K)=SQRTF(VAR(K))
      XBAR(K)=XSUM(K)/XN(K)
      IF(SSUMX(K,I))149,149,405
  405 DCA(K,I)=SUMXT(K,I)/SSUMX(K,I)
      SUMDA(K)=SUMDA(K)+DCA(K,I)*THA(I)
      D(K)=SUMDA(K)/A
      PCA(K,I)=SSUMXT(K,I)/SSUMX(K,I)
      SUMPTA(K)=SUMPTA(K)+PCA(K,I)*THA(I)
      TL(K)=SUMPTA(K)/A
  149 CONTINUE
  150 CONTINUE
      PRINT 200
C     OUTPUT TABLE HEADINGS AND VALUES OF VARIABLES FOR ALL
C     STORMS.
      PRINT 8, (NWSHD(J),J=1,4)
      PRINT 9
      IJI=II+JJ
      DO 152 K=1,NS
      DO 151 I=1,IJI
  151 SUMAP(K)=SUMAP(K)+SUMX(K,I)*THAN(I)
  152 CONTINUE
      DO 157 K=1,NS
      IF(K-KKK(K))156,153,156
  153 IF(ZZ)155,154,155
  154 PRINT 10, NMOP(K),NDAP(K),NYRP(K),NONN,NOON,NOO,NOOO
      GO TO 157
  155 PRINT 11, NMOP(K),NDAP(K),NYRP(K),NONN,NOON,NOO,NOOO
      GO TO 157
  156 PT(K)=SUMAP(K)/A
     OPRINT 12, NMOP(K),NDAP(K),NYRP(K),XBAR(K),VAR(K),
     1STDEV(K),D(K),TL(K),PT(K)
  157 CONTINUE
  158 READ 13, NSTANO
      IF(NSTANO)158, 14,158
  159 CALL EXIT
      END
```

```
C                    APPENDIX E
C
C              CORRELATION COMPUTER PROGRAM
C                 FROM COOLEY AND LOHNES
C      INPUT IS ACTUAL RAW OR TRANSFORMED DATA  TO BE READ IN BY
C      ROWS FROM A MATRIX WITH DIFFERENT TRIALS AS ROWS AND VAR-
C      IABLES AS COLUMNS.  ALL VALUES FOR EACH TRIAL (ONE ROW OF
C      THE DATA MATRIX) ARE READ AND TREATED BEFORE THE NEXT ROW
C      IS READ.  THIS ELIMINATES STORAGE OF THE WHOLE DATA MAT-
C      RIX.  INPUT AND OUTPUT FORMAT STATEMENTS MUST BE CHANGED
C      TO ACCOMODATE THE DATA.  THE FOLLOWING VARIABLE NAMES ARE
C      USED IN THE PROGRAM.  THE FIRST FOUR MUST BE READ IN ON
C      A CARD IMMEDIATELY PRECEEDING THE DATA AND COMPLYING
C      WITH THE FIRST FORMAT.  THE REST DEFINE THE OUTPUT.
C            M=THE NUMBER OF VARIABLES TO BE CORRELATED WITH
C               EACH OTHER.
C            N=THE NUMBER OF OBSERVATIONS OF ALL M VARIABLES.
C       LPUNCH=0 IF CORRELATION MATRIX IS TO BE PUNCHED.
C              =1 IF CORRELATION IS TO BE PRINTED ONLY.
C       LPRINT=0 IF ALL COMPUTED PARAMETERS AND MATRICES ARE TO
C               BE PRINTED AND PUNCHED.
C              =1 IF ONLY CORRELATION MATRIX IS TO BE PRINTED.
C         SX(I)= THE SUM OF ALL VALUES OF A VARIABLE WHERE I IS
C                TAKEN FROM 1 TO M.
C          X(I)= A VALUE OF A VARIABLE WHERE I IS THE SUBSCRIPT
C                OF THE PARTICULAR VARIABLE.
C       SS(I,J)= A MATRIX ENTRY WHICH IS EITHER A SUM OF THE
C                SQUARES OF ALL VALUES OF A VARIABLE OR A SUM OF
C                THE CROSS-PRODUCTS OF ALL VALUES OF TWO VARIAB-
C                LES.
C       SSD(I,J)=A MATRIX ENTRY IN THE DEVIATION SUMS OF SQUARES
C                MATRIX, COMPUTED FROM THE VALUES IN THE
C                PREVIOUS MATRIX.
C        D(I,J)= A MATRIX ENTRY IN THE VARIANCE-COVARIANCE MATRIX
C                COMPUTED AS SSD(I,J)/N-1.
C         SD(I)= A STANDARD DEVIATION COMPUTED AS THE SQUARE ROOT
C                OF THE VARIANCES,D(I,I), I=1,M
C        R(I,J)= A CORRELATION COEFFICIENT IN THE CORRELATION
C                MATRIX.
C       OUTPUT FROM THE PROGRAM IS LABELLED THROUGH OUTPUT FORMAT
C       STATEMENTS.
      ODIMENSION X(32),SX(32),SS(32,32),SSD(32,32),D(32,32),
      1R(32,32),SD(32),XM(32)
       COMMON X,M
       READ 1, M, N, LPUNCH, LPRINT
     1 FORMAT (I2,I2,I1,I1)
       NTRIAL=N
```

```
C     COMPUTE SUMS, SUMS OF SQUARES, SUMS OF CROSS-PRODUCTS
      DO 2 I=1,M
      SX(I)=0.0
      DO 2 J=I,M
    2 SS(I,J)=0.0
    3 READ 4, (X(I),I=1,M)
    4 FORMAT ((10X,10F7.0))
      CALL TRANFO
      DO 5 I=1,M
      SX(I)=SX(I)+X(I)
      DO 5 J=I,M
    5 SS(I,J)=SS(I,J)+X(I)*X(J)
      NTRIAL=NTRIAL-1
      XN=N
C     SET LOWER TRIAGLE EQUAL TO UPPER TRIANGLE AND COMPUTE THE
C           DEVIATION SUMS OF SQUARES AND CROSS-PRODUCTS MATRIX
      IF (NTRIAL) 6,6,3
    6 DO 7 I=1,M
      DO 7 J=I,M
      SSD(I,J)=(XN*SS(I,J)-SX(I)*SX(J))/XN
      SS(J,I)=SS(I,J)
    7 SSD(J,I)=SSD(I,J)
C     COMPUTE  STANDARD DEVIATIONS
      DO 8 I=1,M
      XM(I)=SX(I)/XN
    8 SD(I)=SQRTF(SSD(I,I)/(XN-1.0))
      PRINT 9
      PUNCH 9
    9 FORMAT ( 1H1,30X,27H CORRELATION PROGRAM OUTPUT /)
C     OPTION TO PRINT NUMBER OF VARIABLES,NUMBER OF OBSERVA-
C     TIONS, MEANS, STANDARD DEVIATIONS, SUMS OF SQUARES,
C     DEVIATION SUMS OF SQUARES
      IF (LPRINT) 10,10,22
   10 PRINT 11, M,N
      PUNCH 11, M,N
   110FORMAT ( 25X, 4H FOR,I3, 15H VARIABLES WITH,I3,
     113H OBSERVATIONS //)
      PRINT 12
      PUNCH 12
   12 FORMAT ( 25X, 37H MEANS IN THE ORDER OF VARIABLE INPUT /)
      PRINT 13,(XM(I),I=1,M)
      PUNCH 13,(XM(I),I=1,M)
   13 FORMAT((14X,6F10.6)/)
      PRINT 14
      PUNCH 14
   14 FORMAT ( 27X, 33H STANDARD DEVIATIONS OF VARIABLES /)
      PRINT 15,(SD(I),I=1,M)
```

```
      PUNCH 15,(SD(I),I=1,M)
   15 FORMAT((14X,6F10.6)/)
      PRINT 16
   16 FORMAT ( 23X, 35H SUMS OF SQUARES AND CROSS PRODUCTS,
     17H MATRIX /)
   17 FORMAT (14X,4H ROW,I3/(14X,6F10.4)/)
      DO 18 I=1,M
   18 PRINT 17,I,(SS(I,J),J=1,M)
      PRINT 19
   190FORMAT ( 18X, 36H DEVIATION SUMS OF SQUARES AND CROSS,
     116H PRODUCTS MATRIX /)
   20 FORMAT (14X,4H ROW,I3/(14X,6F10.4)/)
      DO 21 I=1,M
   21 PRINT 20,I,(SSD(I,J),J=1,M)
C     COMPUTE VARIANCE-COVARIANCE MATRIX
   22 DO 23 I=1,M
      DO 23 J=I,M
      D(I,J)=SSD(I,J)/(XN-1.0)
   23 D(J,I)=D(I,J)
C     OPTION TO PRINT VARIANCE-COVARIANCE MATRIX
      IF (LPRINT) 24,24,28
   24 PRINT 25
      PUNCH 25
   25 FORMAT ( 30X, 27H VARIANCE-COVARIANCE MATRIX /)
   26 FORMAT (14X,4H ROW,I3/(14X,6F10.5)/)
      DO 27 I=1,M
      PUNCH 26,I,(D(I,J),J=1,M)
   27 PRINT 26,I,(D(I,J),J=1,M)
C     COMPUTE CORRELATION MATRIX
   28 DO 29 I=1,M
      DO 29 J=I,M
      R(I,J)=D(I,J)/(SD(I)*SD(J))
   29 R(J,I)=R(I,J)
C     OPTION TO PUNCH CORRELATION MATRIX
      PRINT 30
      PUNCH 30
   30 FORMAT ( 34X, 19H CORRELATION MATRIX /)
   31 FORMAT (14X,4H ROW,I3/(14X,6F10.6)/)
      DO 32 I=1,M
   32 PRINT 31,I,(R(I,J),J=1,M)
      IF (LPUNCH)33,33,35
   33 DO 34 I=1,M
   34 PUNCH 31,I,(R(I,J),J=1,M)
   35 CALL EXIT
      END
```

```
      SUBROUTINE TRANFO
      DIMENSION X(32)
      COMMON X,M
      DO 5 I=1,M
      IF(I-28)2,1,2
    1 X(I)=X(I)+32.0
    2 IF(X(I))4,3,4
    3 X(I)=.000001
    4 X(I)=ALOG10(X(I))
    5 CONTINUE
      RETURN
      END
```

```
C                          APPENDIX F
C
C                 PRINCIPAL COMPONENT PROGRAM
C                   FROM COOLEY AND LOHNES
C          THIS PROGRAM PERFORMS PRINCIPAL-COMPONENT ANALYSIS
C       THAT BEGINS WITH EITHER THE CORRELATION OR DISPERSION
C       MATRIX, RATHER THAN RAW SCORES. MATRICES UP TO 60TH
C      ORDER ARE POSSIBLE, BUT THE DIMENSION STATEMENT COULD
C      EASILY BE MODIFIED IF LARGER MATRICES WERE NECESSARY.
C
C       INPUT.
C            CONTROL CARD 1 CONTAINS (COL 1-48) = PROBLEM IDEN-
C      TIFICATION WHICH MAY BE ALPHABETIC, (COL 49-50)=SIZE OF
C      MATRIX(RIGHT-JUSTIFIED) IF M=0, THEN CALL EXIT.  THE
C      MATRIX FOLLOWS CARD 1 AND IS READ IN BY A SUBROUTINE
C      WHICH MUST BE STORED TO THE RIGHT OF THE MAIN DIAGONAL
C      OF R.  SUBROUTINE HDIAG IS REQUIRED.
C
C       OUTPUT.
C           PRINTED OUTPUT INCLUDES THE ROOTS OF MATRIX, NORM-
C      ALIZED VECTORS, PERCENTAGE OF TRACE ACCOUNTED FOR BY THE
C      ROOTS, AND THE FACTOR LOADINGS.
       DIMENSION PROB (12)
       DIMENSION R(60,60),SS(60,60),IQ(60),FRACT(60),X(60)
       INTEGER PROB
C      READING IN THE CONTROL CARD
  100  READ (105,121) PROB,M
  121  FORMAT(12A4,I2)
       IF(M)  1000,1000,210
  210  WRITE (108,21)PROB
       WRITE (106,21)PROB
   21  FORMAT(1H1,24X,12A4///)
       WRITE(108,25)
       WRITE(106,25)
   25  FORMAT(30X,28H ORIGINAL CORRELATION MATRIX/)
       CALL RHDIAG(R,M)
       T = M
       CALL HDIAG (R,M,0,SS,NR)
        WRITE(108,710) NR
        WRITE(106,710) NR
 7100FORMAT(1H0,22X,31H NO. OF ROTATIONS FROM JACOBIAN,
     113H SUBROUTINE =,I3//)
       WRITE (108,1)
       WRITE (106,1)
     1 FORMAT (22X,36H CHARACTERISTIC ROOTS OF CORRELATION,
     17H MATRIX/)
       WRITE (108,3) (R(I,I), I=1,M)
```

```
      WRITE (106,3) (R(I,I), I=1,M)
    3 FORMAT((14X,6F10.6))
      WRITE (108,4)
      WRITE (106,4)
    4OFORMAT(1H0,32X,24H NORMALIZED EIGENVECTORS/25X,
     138H NOTE THAT VECTORS ARE WRITTEN IN ROWS/)
    5 FORMAT((14X,6F10.4))
      DO 6  J=1,M
      WRITE (106,7) J,(SS(I,J), I = 1,M)
    6 WRITE (108,7) J,(SS(I,J), I = 1,M)
    7 FORMAT(14X,7H VECTOR,I3/(14X,6F10.6))
      DO 10  I=1,M
   10 FRACT(I)= (R(I,I) / T)*100.0
      WRITE (108,11)
      WRITE (106,11)
   11OFORMAT(1H0,15X,37H PERCENTAGE OF VARIANCE ACCOUNTED FOR,
     120H BY EACH EIGENVECTOR /)
      WRITE (108,5)(FRACT(I), I=1,M)
      WRITE (106,5)(FRACT(I), I=1,M)
      DO 111 I =2,M
  111 FRACT (I) = FRACT (I) + FRACT (I-1)
      WRITE (108,112)
      WRITE (106,112)
  112 FORMAT(1H0,22X,23H ACCUMULATED PERCENTAGE/)
      WRITE (108,5)(FRACT (I), I = 1,M)
      WRITE (106,5)(FRACT (I), I = 1,M)
      WRITE (108,113)
      WRITE (106,113)
  113OFORMAT(1H0,37X,14H FACTOR MATRIX/25X,10H NOTE THAT,
     128H FACTORS ARE WRITTEN IN ROWS /)
      DO 12  I = 1,M
   12 R(I,I) = SQRT(R(I,I))
      DO 13  J = 1,M
      DO 13  I = 1,M
   13 SS(I,J) = SS(I,J) * R(J,J)
      DO 14  J = 1,M
      WRITE (106,15) J,(SS(I,J), I =1,M)
   14 WRITE (108,15) J,(SS(I,J), I =1,M)
   15 FORMAT(14X,7H FACTOR,I3/(14X,6F10.6))
      GO TO 100
 1000 CALL EXIT
      END
```

```
      SUBROUTINE RHDIAG(R,M)
      DIMENSION R(60,60)
    1 FORMAT (14X,4H ROW,I3/(14X,6F10.6))
    2 FORMAT((14X,6F10.6))
      DO 3 I=1,M
      READ 2, (R(I,J),J=1,M)
    3 PRINT 1, I, (R(I,J),J=1,M)
      RETURN
      END
C     SUBROUTINE HDIAG.
C
C          PROGRAMMED BY F.J. CORBATO AND M. MERWIN OF THE M.
C     I.T. COMPUTATION CENTER.
C          THIS SUBROUTINE COMPUTES THE EIGENVALUES AND EIGEN-
C     VECTORS OF A REAL SYMMETRIC MATRIX, H, OF ORDER N (WHERE
C     N MUST BE LESS THAN 61), AND PLACES THE EIGENVALUES IN
C     THE DIAGONAL ELEMENTS OF THE MATRIX H, AND PLACES THE
C     EIGENVECTORS (NORMALIZED) IN THE COLUMNS OF THE MATRIX U
C     IEGEN IS SET AS 1 IF ONLY EIGENVALUES ARE DESIRED, AND
C     IS SET TO 0 WHEN VECTORS ARE REQUIRED.  NR CONTAINS THE
C     NUMBER OF ROTATIONS DONE.
C
      SUBROUTINE HDIAG (H,N,IEGEN,U,NR)
      DIMENSION H(60,60),U(60,60),X(60),IQ(60)
      IF(IEGEN)  15,10,15
   10 DO 14 I=1,N
      DO 14 J=1,N
      IF(I-J)  12,11,12
   11 U(I,J)=1.0
      GO TO 14
   12 U(I,J)=0.
   14 CONTINUE
   15 NR = 0
      IF (N-1)  1000,1000,17
C     SCAN FOR LARGEST OFF-DIAGONAL ELEMENT IN EACH ROW
C     X(I) CONTAINS LARGEST ELEMENT IN ITH ROW
C     IQ(I) HOLDS SECOND SUBSCRIPT DEFINING POSITION OF ELEMENT
   17 NMI1=N-1
      DO 30 I=1,NMI1
      X(I) = 0.
      IPL1=I+1
      DO 30 J=IPL1,N
      IF( X(I) - ABS( H(I,J)))  20,20,30
   20 X(I)=ABS(H(I,J))
      IQ(I)=J
   30 CONTINUE
C     SET INDICATOR FOR SHUT-OFF.RAP=2**-27,NR=NO. OF ROTATIONS
```

```
      RAP=7.450580596E-9
      HDTEST=1.0E38
C     FIND MAXIMUM OF X(I) S FOR PIVOT ELEMENT AND
C     TEST FOR END OF PROBLEM
   40 DO 70 I=1,NMI1
      IF (I-1)  60,60,45
   45 IF ( XMAX- X(I))  60,70,70
   60 XMAX=X(I)
      IPIV=I
      JPIV=IQ(I)
   70 CONTINUE
C     IS MAX. X(I) EQUAL TO ZERO, IF LESS THAN HDTEST, REVISE
C     HDTEST
      IF ( XMAX)  1000,1000,80
   80 IF (HDTEST)  90,90,85
   85 IF (XMAX - HDTEST)  90,90,148
   90 HDIMIN = ABS( H(1,1) )
      DO 110  I= 2,N
      IF (HDIMIN- ABS( H(I,I))) 110,110,100
  100 HDIMIN=ABS(H(I,I))
  110 CONTINUE
      HDTEST=HDIMIN*RAP
C     RETURN IF MAX.H(I,J)LESS THAN(2**-27)ABS(H(K,K)-MIN)
      IF (HDTEST- XMAX) 148,1000,1000
  148 NR = NR+1
C     COMPUTE TANGENT, SINE AND COSINE,H(I,I),H(J,J)
  150 TANG=SIGN(2.0,(H(IPIV,IPIV)-H(JPIV,JPIV)))*H(IPIV,JPIV)/
     1(ABS(H(IPIV,IPIV)-H(JPIV,JPIV))+SQRT((H(IPIV,IPIV)-H(
     2JPIV,JPIV))**2+4.0*H(IPIV,JPIV)**2))
      COSINE=1.0/SQRT(1.0+TANG**2)
      SINE=TANG*COSINE
      HII=H(IPIV,IPIV)
      H(IPIV,IPIV)=COSINE**2*(HII+TANG*(2.*H(IPIV,JPIV)+TANG*H
     1(JPIV,JPIV)))
      H(JPIV,JPIV)=COSINE**2*(H(JPIV,JPIV)-TANG*(2.*H(IPIV,JPI
     1V)-TANG*HII))
      H(IPIV,JPIV)=0.
C     PSEUDO RANK THE EIGENVALUES
C     ADJUST SINE AND COS FOR COMPUTATION OF H(IK) AND U(IK)
      IF ( H(IPIV,IPIV) - H(JPIV,JPIV)) 152,153,153
  152 HTEMP = H(IPIV,IPIV)
      H(IPIV,IPIV) = H(JPIV,JPIV)
      H(JPIV,JPIV) = HTEMP
C     RECOMPUTE SINE AND COS
      HTEMP = SIGN (1.0,-SINE) * COSINE
      COSINE = ABS (SINE)
      SINE = HTEMP
```

```
  153 CONTINUE
C     INSPECT THE ROS BETWEEN I+1 AND N-1 TO DETERMINE
C     WHETHER A NEW MAXIMUM VALUE SHOULD BE COMPUTED SINCE
C     THE PRESENT MAXIMUM IS IN THE I OR J ROW.
      DO 350 I=1,NMI1
      IF(I-IPIV)  210,350,200
  200 IF(I-JPIV)  210,350,210
  210 IF (IQ(I)-IPIV)  230,240,230
  230 IF (IQ(I)-JPIV)  350,240,350
  240 K=IQ(I)
  250 HTEMP=H(I,K)
      H(I,K)=0.
      IPL1=I+1
      X(I) =0.
C     SEARCH IN DEPLETED ROW FOR NEW MAXIMUM
      DO 320 J=IPL1,N
      IF ( X(I)-ABS( H(I,J)) ) 300,300,320
  300 X(I) = ABS(H(I,J))
      IQ(I)=J
  320 CONTINUE
      H(I,K)=HTEMP
  350 CONTINUE
      X(IPIV) =0.
      X(JPIV) =0.
C     CHANGE THE OTHER ELEMENTS OF H
      DO 530 I=1,N
      IF(I-IPIV)  370,530,420
  370 HTEMP = H(I,IPIV)
      H(I,IPIV) = COSINE*HTEMP + SINE*H(I,JPIV)
      IF ( X(I) - ABS( H(I,IPIV)) ) 380,390,390
  380 X(I) = ABS(H(I,IPIV))
      IQ(I) = IPIV
  390 H(I,JPIV) = -SINE*HTEMP + COSINE*H(I,JPIV)
      IF ( X(I) - ABS( H(I,JPIV)) ) 400,530,530
  400 X(I) = ABS(H(I,JPIV))
      IQ(I) = JPIV
      GO TO 530
  420 IF (I-JPIV) 430,530,480
  430 HTEMP = H(IPIV,I)
      H(IPIV,I) = COSINE*HTEMP + SINE*H(I,JPIV)
      IF ( X(IPIV) - ABS( H(IPIV,I)) ) 440,450,450
  440 X(IPIV) = ABS(H(IPIV,I))
      IQ(IPIV)=I
  450 H(I,JPIV) = -SINE*HTEMP + COSINE*H(I,JPIV)
      IF ( X(I) - ABS( H(I,JPIV)) ) 400,530,530
  480 HTEMP = H(IPIV,I)
      H(IPIV,I) = COSINE*HTEMP + SINE*H(JPIV,I)
```

```
      IF ( X(IPIV) - ABS( H(IPIV,I)) ) 490,500,500
  490 X(IPIV) = ABS(H(IPIV,I))
      IQ(IPIV) = I
  500 H(JPIV,I) = -SINE*HTEMP + COSINE*H(JPIV,I)
      IF ( X(JPIV) - ABS( H(JPIV,I)) ) 510,530,530
  510 X(JPIV) = ABS(H(JPIV,I))
      IQ(JPIV) = I
  530 CONTINUE
C     TEST FOR COMPUTATION OF EIGENVECTORS
      IF(IEGEN) 40,540,40
  540 DO 550 I=1,N
      HTEMP=U(I,IPIV)
      U(I,IPIV)=COSINE*HTEMP+SINE*U(I,JPIV)
  550 U(I,JPIV)=-SINE*HTEMP+COSINE*U(I,JPIV)
      GO TO 40
 1000 RETURN
      END
```

```
C                          APPENDIX G
C
C                  VARIMAX ROTATION PROGRAM
C                   FROM COOLEY AND LOHNES
C        THIS PROGRAM WILL PERFORM VARIMAX ROTATION FOR UP TO 75
C      FACTORS AND 100 VARIABLES.
C
C      INPUT.
C      ENTER NEIG (COLS 59 - 60).
C           CARD 1 CONTAINS PROB(C.C. 1-48)= PROBLEM IDENTIFI-
C      CATION WHICH MAY BE ALPHABETIC, N (COL 1-48)= NO. OF
C      VARIABLES, IF N=0, CALL EXIT, L (COLS 52-53) = NO. OF
C      FACTORS, NVALUE (COLS 54-55) = NO. OF ITERATION CYCLES
C      (USE 00 IF THIS TEST IS NOT DESIRED) JTEST (COL 56-57) =
C      NO. OF CYCLES NEW VARIANCE REQUIRED TO EQUAL OLD VARI-
C      ANCE, RORC (COL58) = 0 IF AN INPUT CARD CONTAINS THE
C      LOADINGS FOR A VARIABLE (MATRIX IS IN CARDS AS ROWS)
C      RORC = 1 IF AN INPUT CARD CONTAINS THE LOADINGS FOR A
C      FACTOR (MATRIX IS IN CARDS AS COLUMNS).
C           IF INPUT DATA IS FROM EIGENVECTORS AND NOT FACTORS,
C      NEIG = 0, OTHERWISE NEIG = 1.  IF NEIG IS 0, ENTER
C      EIGENVALUES IN THE ORDER OF RESPECTIVE EIGENVECTORS ON
C      CARD(S) WHICH FOLLOW THE ABOVE DATA.
C          LOADINGS FOLLOW CARD 1.
      DIMENSION A( 50,20),V(50),TV(50),H( 50),AX( 50,20)
     1,HN( 50),HO( 50),PROB(12),EIGVAL(20),AY( 50,20)
      COMMON A,V,TV,NV,N,L,FN,T,B,P
 1001 FORMAT(17HONO.      VARIANCE)
 1002 FORMAT(1H I3,F20.8)
 1003 FORMAT(40X,16HO  FACTOR MATRIX)
 1004 FORMAT(9HOVARIABLEI4/)
 1005 FORMAT(1H 10F11.4)
 1006 FORMAT(12A4,I3,3I2,I1,I2)
 1007 FORMAT(5F14.7)
 1008 FORMAT(27HOOLD H2  NEW H2  DIFFERENCE/)
 1009 FORMAT(F6.3,F8.3,F12.3,I8)
 1010 FORMAT((10X,7F10.6))
 10120FORMAT(1H1,26X,37H OUTPUT FROM VARIMAX ROTATION PROGRAM/
     1/20X,12A4//)
 10130FORMAT(21X,37H EIGENVALUES FROM PRINCIPAL COMPONENT,
     19H ANALYSIS //(14X,6F10.6))
 10140FORMAT(1HO/ 14X,35H NO. ROTATIONS REQUIRED TO MAXIMIZE,
     120H VARIMAX CRITERION = ,I3)
 1015 FORMAT(12A6)
 1016 FORMAT(1HO/33X,22H ORIGINAL EIGENVECTORS/)
 1017 FORMAT(14X,7H VECTOR,I3/(14X,6F10.6))
 1018 FORMAT(1HO/33X,23H ORIGINAL FACTOR MATRIX/)
```

```
1019 FORMAT(14X,7H FACTOR,I3/(14X,6F10.6))
1020 FORMAT(1H0,30X,28H FINAL ROTATED FACTOR MATRIX/)
1021 FORMAT(1H0,32X,24H VARIANCE OF EACH FACTOR//(14X,6F10.6))
10220FORMAT(1H0,17X,37H PERCENTAGE OF VARIANCE ACCOUNTED FOR,
    115H BY EACH FACTOR//(14X,6F10.5))
1023 FORMAT(1H0,22X,23H ACCUMULATED PERCENTAGE//(14X,6F10.4))
10240FORMAT(1H0,29X,30H REDUCED FINAL ROTATED FACTORS/14X,
    19H VARIABLE, 21X,8H FACTORS/)
1025 FORMAT(18X,I2,3X,10F5.2)
1026 FORMAT(/14X,  9HVARIANCES,   10F5.1)
1027 FORMAT((14X,6F10.6))
  24 READ(105,1006)(PROB(I),I=1,12),N,L,NVALUE,JTEST,NRORC,NEIG
     RORC = NRORC
     IF(N)  6000,6000,25
  25 WRITE (108,1012),PROB
     WRITE (106,1012),PROB
     IF (RORC)  26,26,28
  26 DO 27 I = 1,N
  27 READ(105,1010)(A(I,J),J=1,L)
     GO TO 29
  28 DO 20  J= 1,L
  20 READ(105,1027)(A(I,J),I=1,N)
  29 IF(NEIG)31,31,30
  31 READ(105,1010)(EIGVAL(J),J=1,L)
     WRITE(108,1013)(EIGVAL(J),J=1,L)
     WRITE(106,1013)(EIGVAL(J),J=1,L)
     DO 32 J=1,L
  32 EIGVAL(J)=SQRT(EIGVAL(J))
     WRITE(108,1016)
     WRITE(106,1016)
     DO 34 J=1,L
     WRITE(106,1017)J,(A(I,J),I=1,N)
  34 WRITE(108,1017)J,(A(I,J),I=1,N)
     DO 33 J=1,L
     DO 33 I=1,N
  33 A(I,J)=EIGVAL(J)*A(I,J)
     WRITE(108,1018)
     WRITE(106,1018)
     DO 35 J=1,L
     WRITE(106,1019)J,(A(I,J),I=1,N)
  35 WRITE(108,1019)J,(A(I,J),I=1,N)
  30 EPS=0.00116
     NC=0
     TV(1)=0.0
     LL=L-1
     NV=1
     FN=N
```

```
      CONS=1.0/SQRT(2.0)
      DO 3 I=1,N
    3 H(I)=0.0
      DO 4 I=1,N
      DO 4 J=1,L
      H(I)=H(I)+A(I,J)*A(I,J)
    4 CONTINUE
      DO 5 I=1,N
      H(I)=SQRT(H(I))
      DO 5 J=1,L
    5 A(I,J)=A(I,J)/H(I)
  222 CALL VARMAX
  102 LCYCLE=NV-2
      DO 6 I=1,N
      DO 6 J=1,L
    6 AX(I,J)=A(I,J)*H(I)
      LV=NV-1
      IF(NV-50)  9,999,999
    9 IF(LV-NVALUE)  10,999,10
   10 IF(JTEST)  13,13,16
   16 IF(ABS(TV(NV)-TV(LV))-0.0000001)  11,11,13
   11 NC=NC+1
   12 IF(NC-JTEST)  13,999,999
   13 DO 500 J=1,LL
      II=J+1
      DO 500 K=II,L
      AA=0.0
      BB=0.0
      CC=0.0
      DD=0.0
      DO 15 I=1,N
      XX=A(I,J)
      YY=A(I,K)
      U=(XX+YY)*(XX-YY)
      W = 2.0*XX*YY
      CC=CC+(U+W)*(U-W)
      DD=DD+2.0*U*W
      AA=AA+U
      BB=BB+W
   15 CONTINUE
      T=DD-2.0*AA*BB/FN
      B=CC-(AA**2-BB**2)/FN
      P=0.25*ATANF(T/B)
      TAN4P=T/B
      IF(T-B)  1041,1433,1042
 1433 IF(T+B-EPS)  500,1043,1043
 1043 COS4T=CONS
```

```
      SIN4T=CONS
      GO TO 5000
 1041 TAN4T=ABS(T)/ABS(B)
      IF(TAN4T-EPS)  8000,1100,1100
 1100 COS4T=1.0/SQRT(1.0+TAN4T**2)
      SIN4T=TAN4T*COS4T
      GO TO 5000
 8000 IF(B)  1150,500,500
 1150 SINP=CONS
      COSP=CONS
      GO TO 1000
 1042 CTN4T=ABS(T)/ABS(B)
      IF(CTN4T-EPS)  9000,1200,1200
 1200 SIN4T=1.0/SQRT(1.0+CTN4T**2)
      COS4T=CTN4T*SIN4T
      GO TO 5000
 9000 COS4T=0.0
      SIN4T=1.0
 5000 COS2T=SQRT((1.0+COS4T)/2.0)
      SIN2T=SIN4T/(2.0*COS2T)
      COST=SQRT((1.0+COS2T)/2.0)
      SINT=SIN2T/(2.0*COST)
      IF(B)1250,1250,1300
 1300 COSP=COST
      SINP=SINT
      GO TO 7000
 1250 COSP=CONS*COST+CONS*SINT
      SINP=ABS(CONS*COST-CONS*SINT)
 7000 IF(T)1400,1400,1000
 1400 SINP=-SINP
 1000 X=COSP
      Y=SINP
      DO 100 I=1,N
      AIJ    =A(I,J)*X+A(I,K)*Y
      AIK    =-A(I,J)*Y+A(I,K)*X
      A(I,J)=AIJ
  100 A(I,K)=AIK
  500 CONTINUE
      GO TO 222
  999 DO 301 I=1,N
  301 HN(I)=0.0
      DO 303 I=1,N
      H(I)=H(I)*H(I)
      DO 302 J=1,L
  302 HN(I)=HN(I)+AX(I,J)*AX(I,J)
  303 HD(I)=HN(I)-H(I)
      WRITE(108,1014)NV
```

```
      WRITE(106,1014)NV
      WRITE(108,1020)
      WRITE(106,1020)
      XN=N
      DO 305 J=1,L
      H(J)=0.0
      WRITE(108,1019)J,(AX(I,J),I=1,N)
      WRITE(106,1019)J,(AX(I,J),I=1,N)
      DO 305 I=1,N
      AY(I,J)=AX(I,J)**2
      H(J)=H(J)+AY(I,J)
      HN(J)=(H(J)*100.)/XN
  305 CONTINUE
      WRITE(108,1021)(H(J),J=1,L)
      WRITE(106,1021)(H(J),J=1,L)
      WRITE(108,1022)(HN(J),J=1,L)
      WRITE(106,1022)(HN(J),J=1,L)
      DO 306 J=2,L
  306 HN(J)=HN(J)+HN(J-1)
      WRITE(108,1023)(HN(J),J=1,L)
      WRITE(106,1023)(HN(J),J=1,L)
      WRITE(108,1020)
      WRITE(106,1020)
      WRITE(108,1024)
      WRITE(106,1024)
      DO 307 I=1,N
      WRITE(106,1025)I,(AX(I,J),J=1,L)
  307 WRITE(108,1025)I,(AX(I,J),J=1,L)
      WRITE(108,1026)(H(J),J=1,L)
      WRITE(106,1026)(H(J),J=1,L)
      GO TO 24
 6000 CALL EXIT
      END
```

```
      SUBROUTINE VARMAX
      DIMENSION A( 50,20),SA(20),SA2(20),V(50),TV(50)
      COMMON A,V,TV,NV,N,L,FN,T,B,P
      SV=0.0
      NV=NV+1
      DO 6 J=1,L
      SA(J)=0.0
      SA2(J)=0.0
    6 CONTINUE
      DO 8 J=1,L
      DO 7 I=1,N
      SA(J)=SA(J)+A(I,J)*A(I,J)
    7 SA2(J)=SA2(J)+(A(I,J)*A(I,J))**2
      SA(J)=SA(J)**2
    8 V(J)=(FN*SA2(J)-SA(J))/FN**2
      DO 9 J=1,L
    9 SV=SV+V(J)
      TV(NV)=SV
      RETURN
      END
```

APPENDIX H

MEANS AND STANDARD DEVIATIONS OF VARIABLES

VARIABLE	RAW DATA		TRANSFORMED DATA	
	MEAN	STD. DEV.	MEAN	STD. DEV.
A	21.846298	16.251465	1.254370	0.251140
SHP	2.981933	0.648277	0.465232	0.088734
AZ	205.619995	57.481766	2.295727	0.125083
ELEV	3496.000000	619.873047	3.536823	0.077241
GNDS	0.044714	0.025749	-1.434831	0.285002
GNDL	0.669614	0.240575	-0.209380	0.187961
FREQ	12.685501	6.949473	1.033136	0.256508
L	12.912598	5.838099	1.072739	0.178872
S	0.007940	0.004395	-2.135564	0.152279
USE	13.959315	20.957428	0.764478	0.584418
INFR	0.096311	0.003343	-1.016577	0.015277
POND	0.018874	0.009896	-2.017157	1.049355
I	0.088012	0.084098	-1.247360	0.426518
ISD	0.103150	0.109601	-1.282492	0.569079
D	51.339890	83.787094	1.326674	0.581200
TDF	3.995995	6.934675	-1.047191	2.702044
TPCP	1.367451	1.546431	-0.091693	0.462670
API	0.354819	0.347613	-0.723012	0.565864
SOLM	14.580000	3.881351	1.147837	0.121640
WDIR	4.179998	1.913325	0.570156	0.223902
WEEK	21.119980	7.272090	1.293948	0.174260
AIRT	46.359482	17.613785	1.619226	0.232972
ATSD	10.277411	4.350557	0.970650	0.201078
WVEL	10.679915	3.823489	1.001662	0.156743
WVSD	6.786855	1.855226	0.816315	0.116398
SOLT	50.308075	17.824463	1.670814	0.171252
STSD	4.549333	2.450234	0.604456	0.219946
DEGD	56.365906	20.872681	1.710973	0.206981
SWEQ	0.245000	0.408353	-3.738943	2.807280
QMAX	147.221939	289.630127	1.781961	0.546014
RUNF	0.179998	0.313125	-1.169482	0.629060

LITERATURE CITED

Anderson, Henry W., "General Report on International Hydrology Symposium," Proceedings, The International Hydrology Symposium, Vol. 2, September, 1967, Fort Collins, Colorado.

Baggaley, A. R., Intermediate Correlational Methods, New York: John Wiley and Sons, Inc., 1964.

Boner, Fred C., Interim Report on the Frequency and Magnitude of Floods in Eastern Montana, U.S.G.S., Helena, Montana, 1963.

Boner, F. C., and R. J. Omang, "Magnitude and Frequency of Floods from Drainage Areas Less Than 100 Square Miles in Montana," U.S.G.S. Open File Report, Helena, Montana, 1967.

Chow, Ven Te, Handbook of Applied Hydrology, A Compendium of Water-Resources Technology, New York: McGraw Hill Book Co., 1964.

Cooley, W. W., and P. R. Lohnes, Multivariate Procedures for the Behavioral Sciences, New York: John Wiley and Sons, 1962.

DuBois, Philip H., Multivariate Correlational Analysis, New York: Harper and Bros., 1957.

Eiselstein, Leo M., "A Principal Component Analysis of Surface Runoff Data from a New Zealand Alpine Watershed," Proceedings, The International Hydrology Symposium, Vol. 1, September, 1967, Fort Collins, Colorado, pp. 479-489.

Fruchter, Benjamin, Introduction to Factor Analysis, New York: D. Van Nostrand Co., 1954.

Guilford, J. P., "When Not to Factor Analyze," Psychology Bulletin, Vol. 49, 1952, pp. 26-37.

Harman, Harry H., Modern Factor Analysis, Chicago Press, 1967.

Harris, B., Sharp, A. L., Gibbs, A. E., and W. J. Owen, "An Improved Statistical Model for Evaluating Parameters Affecting Water Yields of River Basins," J. Goephys. Res., Vol. 66, 1961, p. 3319.

Holzinger, K. J., and H. H. Harman, Factor Analysis, Chicago: Univ. of Chicago Press, 1941.

Hotelling, Harold, "Analysis of a Complex of Statistical Variables into Principal Components," J. Educational Psychology, Vol. 24, 1933, pp. 417-441 and 498-520.

Kaiser, H. F., "The Varimax Criterion for Analytic Rotation in Factor Analysis," Psychometrika, Vol. 23, 1958, pp. 187-200.

Kaiser, H. F., "Computer Program for Varimax Rotation in Factor Analysis," Educational and Psychological Measurement, Vol. 19, 1959, pp. 413-420.

Kendall, M. G., A Course in Multivariate Analysis, New York: Hafner Publishing Co., 1957.

Matalas, N. C., and Barbara J. Reiher, "Some Comments on the Use of Factor Analysis," Water Resources Res., Vol. 3, No. 1, 1967, pp. 213-224.

Ralston, A., and H. S. Wilf, Mathematical Methods for Digital Computers, New York: John Wiley and Sons, 1960.

Rice, Raymond M., "Multivariate Methods Useful in Hydrology," Proceedings, The International Hydrology Symposium, Vol. 1, September, 1967, Fort Collins, Colorado, pp. 471-478.

Sharp, A. L., and A. K. Biswas, "Research Needs in Surface-Water Hydrology," Journal of Hydraulics Division, ASCE, Vol. 91, No. HY5, Proc. Paper 4464, September, 1965, p. 282.

Sharp, A. L., Gibbs, A. E., Owen, W. J., and B. Harris, "Application of the Multiple Regression Approach in Evaluating Parameters Affecting Water Yields of River Basins," J. Geophys. Res., Vol. 65, 1960, pp. 1273-1286.

Snyder, Willard M., "Some Possibilities for Multivariate Analysis in Hydrologic Studies," J. Geophys. Res., Vol. 67, 1962, p. 721.

Thurstone, L. L., Multiple Factor Analysis, Chicago: Univ. of Chicago Press, 1947.

Wallis, J. R., "Multivariate Statistical Methods in Hydrology - - a comparison using data of known functional relationship," Water Resources Res., Vol. 1, No. 4, 1965, pp. 447-461.

Wallis, J. R., "Factor Analysis in Hydrology - An Agnostic View," Water Resources Res., Vol. 4, No. 3, June, 1968, pp. 521-527.

Williams, T. T., "Selection of Small Watersheds for Hydrologic Study," Drainage Correlation Research Project Interim Report # 1, Mont. Highway Comm., 1965.

Williams, T. T., "Precipitation and Streamflow Instrumentation for Small Watersheds," Drainage Correlation Research Project Interim Report # 2, Mont. Highway Comm., 1965.

Wong, Shue Tuck, "A Multivariate Statistical Model for Predicting Mean Annual Flood in New England," Annals of the Association of American Geographers, Vol. 53, No. 3, 1963, pp. 298-311.